

DOCUMENT RESUME

ED 238 894

TM 830 628

AUTHOR Lewy, A.; Kugelmass, S.
TITLE Decision Oriented Evaluation in Education: The Case of Israel.
REPORT NO ISBN-0-86689-013-0
PUB DATE 81
NOTE 217p.; Papers presented at the joint Israel-American Seminar on Educational Evaluation (Jerusalem, Israel, June, 1980).
AVAILABLE FROM International Science Services, 2242 Mt. Carmel Ave., Glenside, PA 19038 (\$20.00).
PUB TYPE Books (010) -- Collected Works - Conference Proceedings (021)
EDRS PRICE MF01/PC09 Plus Postage.
DESCRIPTORS *Curriculum Development; *Decision Making; Educational Policy; *Evaluation; Evaluation Methods; *Evaluation Utilization; Evaluators; Foreign Countries; *International Educational Exchange; Models; Research Methodology
IDENTIFIERS *Israel

ABSTRACT

This book contains a series of papers presented at an Israel-American seminar on Educational Evaluation. Sol Kugelmass describes strategies employed to increase the utilization of evaluation results in the process of formulating educational policy. Nachum Blass illustrates the relationship between evaluation and policymaking by discussing an evaluation study designed to examine the outcomes of the Educational Reform in Israel. Evaluation problems of the Israeli matriculation examinations are discussed by Baruch Nevo and Gershon Ben-Shakhar. Arie Lewy describes the Israel Curriculum Center organizational framework and elaborates on the communication pattern of the development/evaluation teams. Pinchas Tamir presents decisions made on the basis of evaluation findings. Daniel Davis argues for standardized methods for evaluating the potential benefits of innovative programs. David Nevo describes a Multi-Method and Multiple Clientele Evaluation Matrix model. Itai Zak examines methodological problems related to dealing with complex data sets. Robert Boruch summarizes the presented works, pointing out similarities and differences between evaluation patterns in the educational systems of the United States and Israel. Boruch places emphasis on communication links established between evaluators and decisionmakers in both countries. (PN)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED238894

DECISION ORIENTED EVALUATION IN EDUCATION

THE CASE OF ISRAEL

A. Lewy and S. Kugelmass

G. Ben Shakhar
N. Blass
R.F. Boruch
D.J. Davis
B. Nevo
D. Nevo
P. Tamir
I. Zak

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

- ☒ This document has been reproduced as received from the person or organization originating it.
- ☐ Minor changes have been made to improve reproduction quality.
- ☐ Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

M. Belaton

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

INTERNATIONAL SCIENCE SERVICES
Philadelphia Rehovot

TM 830 628

DECISION ORIENTED EVALUATION IN EDUCATION

THE CASE OF ISRAEL

A. Lewy and S. Kugelmass

G. Ben Shakhar

N. Blass

R.F. Boruch

D.J. Davis

B. Nevo

D. Nevo

P. Tamir

I. Zak

INTERNATIONAL SCIENCE SERVICES

Philadelphia

Rehovot

International Science Services
2242 Mt. Carmel Avenue
Glenside, Pa. 19038, U.S.A.
or
POB 2039, Rehovot, Israel

Editorial assistance and English style by Dr. Shmuel Himelstein

Copyright © 1981 by Balaban International Science Services

Library of Congress Cataloging in Publication Data
Main entry under title:

Decision oriented evaluation in education.

"Papers... presented at the Israel-U.S.A.
Seminar on Evaluation (held) in Jerusalem (Israel) in
June 1980" - Introd.

Includes bibliographies and index.

1. Educational surveys-Israel-Congresses.
 2. Educational surveys-United States-Congresses.
 3. Educational research-Israel-Congresses.
 4. Educational research-United States-Congresses.
- I. Lewy, Arie, 1923- . II. Kugelmass, Solomon,
1926- . III. Israel-U.S.A. Seminar on Evalua-
tion (1980 : Jerusalem)

LB2823.D43 370'.7'8
ISBN 0-86689-013-0

81-10800
AACR2

All rights reserved.

No part of this book may be reproduced by any means, nor transmitted into
a machine language without the written permission of the publisher.

CONTENTS

Contributors	vi
Foreword	vii
Preface	ix
Introduction	xi
On Evaluation Policy in the United States and Israel <u>Robert F. Boruch</u>	1
Considerations Toward a Policy of Evaluation Research: The Case of the Chief Scientist in the Israeli Ministry of Education <u>Sol Kugelmass</u>	33
Research and Policy-Making in the Israeli Educational System - a Case Study <u>Nachum Blass</u>	51
Some Thoughts Regarding the Possibility of Evaluating a National Matriculation Examinations System <u>Gershon Ben-Shakhar and Baruch Nevo</u>	97
Sharing Responsibilities for Evaluation: Cooperation between the Development and Evaluation Teams <u>Arieh Lewy</u>	107
Continuous Evaluation as an Integral Part of Curriculum Development: The Evaluation of the High School Biology Project at the Israel Science Teaching Center <u>Pinchas Tamir</u>	133
Standardized Evaluation of Education Programs <u>Daniel J. Davis</u>	155
The Evaluation of a Multi-Dimensional Project <u>David Nevo</u>	169
Latent Variables in Causal Models: An Example from Educational Research <u>Itai Zak</u>	195

CONTRIBUTORS

ROBERT F. BORUCH, Ph D; Professor of Psychology and Education, Northwestern University (Evanston, Illinois); Past President, Council for Applied Social Research; member, Board of Trustee, Evaluation Research Society; chairman and member of numerous committees of the National Academy of Sciences and the Social Science Research Council.

NACHUM BLASS, M A; Research Coordinator, Office of the Chief Scientist, Ministry of Education (Israel).

DANIEL J. DAVIS, Ph D; Senior Lecturer, School of Education, The Hebrew University of Jerusalem (Israel).

BARUCH NEVO, Ph D; Senior Lecturer, Department of Psychology, Haifa University (Israel).

DAVID NEVO, Ph D; Senior Lecturer of Educational Evaluation, School of Education, Tel Aviv University (Israel).

SOL KUGELMASS, Ph D; Chief Scientist, Ministry of Education, Professor, Department of Psychology, Hebrew University of Jerusalem (Israel).

ARIEH LEWY, Ph D; Associate Professor, School of Education, Tel Aviv University (Israel); Head, evaluation unit, Israel Curriculum Center; bureau member, International Association for the Evaluation of Educational Achievement; editor, *Studies in Educational Evaluation*.

PINCHAS TAMIR, Ph D; Associate Professor, School of Education and Israel Science Teaching Center, The Hebrew University of Jerusalem (Israel).

ITAI ZAK, Ph D; Senior Lecturer; Director, Evaluation Research and Measurement Unit, School of Education, Tel Aviv University; Chairman, Israeli Educational Research Association (Israel).

GERSHON BEN-SHAKHAR, Ph D; Lecturer, Department of Psychology, The Hebrew University of Jerusalem (Israel).

FOREWORD

This book contains proceedings of a joint Israel-American seminar on Educational Evaluation. The cooperation between Israeli and American researchers in the field of education has a long history. Numerous Israeli researchers received their training in the United States and are graduates of American universities. American professors frequently spend their sabbaticals at Israeli universities. American journals of educational research publish many articles of Israeli researchers and, on the other hand, books and articles published in Israel are full of references to studies carried out in the United States.

During several decades, spontaneous cooperation took place in the domain of educational research between Israeli and American scientists. Such spontaneous cooperation served as a basis for formally approved and institutionalized cooperation, which culminated in a cultural agreement between the two countries. Signed in Jerusalem in 1978, the U.S. - Israel Memorandum of Understanding on Education includes provisions for educational colloquia to take place alternately in Washington and Jerusalem, summer seminars for curriculum planners, along with this research seminar.

The Ministry of Education and Culture attributes great importance to research and development activities of the research community in Israel, and in the past as well as presently it has allocated funds to support scientific activities in Israel. The establishment of the Chief Scientist's Office in the Ministry reflects a sincere attempt to increase utilization of research in monitoring the activities of the educational system.

I wish to thank Dr. Robert Boruch of Northwestern University, Evanston, Illinois, for participating in the seminar. His analytical comparison of the links between evaluation and decision-making in the educational systems of Israel and the United States will certainly contribute to the efficient utilization of evaluation results in Israel. I also wish to express my appreciation to the Israeli participants of the seminar; evaluation experts from almost all the institutions of higher education in this country contributed to the seminar. And finally, my special thanks to all those who devoted their time and energies to the publishing of the proceedings of the seminar.

Eliezer Shmueli
Director General
Ministry of Education and Culture

PREFACE

The origins of this book lie in a seminar conducted in accordance with the U.S. - Israel Memorandum of Understanding of November 1978, signed by Mr. Joseph Califano, then Secretary of Education, U.S. Department of Education, and Mr. Zevulun Hammer, Minister of Education and Culture of Israel. The seminar which resulted in this book was one of a series which brought together professionals involved in educational evaluation from the United States and Israel.

The authors would like to take this opportunity to express their appreciation for all those who helped in facilitating the seminar and in producing this book. Special thanks are due to:

The U.S. Department of Education, for its involvement in the venture;

The U.S. National Institute of Education, which funded the seminars and subsequent publications which came out of them, and the U.S.-Israel Educational Foundation which administered the program in Israel. Special thanks are due to Mr. Daniel M. Krauskopf, Executive Secretary of the U.S.-Israel Educational Foundation, for his invaluable assistance in seeing to the many details involved in ensuring the success of this enterprise.

INTRODUCTION

This book contains a series of papers which were presented at the Israel-U.S.A. Seminar on Evaluation in Jerusalem (Israel) in June 1980. The evaluation seminar was one of five cooperative Israel-U.S.A. seminars which were held in Israel in 1980 in accordance with the Israel-U.S.A. cultural agreement signed by the Ministry of Education and Culture of Israel and the U.S. Department of Education.

The Israeli participants in the seminar were from the Hebrew University, Jerusalem, the Tel Aviv University, the University of Haifa, the Technion and the Office of the Chief Scientist of the Ministry of Education and Culture in Israel.

The papers presented by the Israeli participants focus on the impact of evaluation activities and decision-making in the field of education. Kugelmass describes the nature of the task of the Office of the Chief Scientist of the Ministry of Education and Culture and elaborates on strategies employed by this office to increase the utilization of evaluation results in the process of formulating educational policy. Blass illustrates the relationship between evaluation and policy-making by discussing the impact of a particular evaluation study which was designed to examine the outcomes of the Educational Reform in Israel. Problems related to the evaluation of the Israeli matriculation examinations are discussed in the article by Baruch Nevo and Gershon Ben-Shakhar.

Two articles deal with activities in the field of curriculum evaluation. Lewy's article describes the organizational framework of evaluation activities in the Israel Curriculum Center (ICC) and elaborates on the communication pattern which has been established between the development teams and the evaluation team. Tamir focuses on the evaluation of a particular curriculum project which was developed by the Israel Science Teaching Center, and presents a list of decisions made on the basis of the findings of a series of evaluation studies.

Two articles deal with the evaluation of large-scale action programs in the Israeli educational system. Davis argues for evaluating the potential benefits of the innovative program, by examining program outcomes under

optimal implementation conditions. David Nevo addresses himself to the problem of providing information to a multiple clientele. The evaluation model described by David Nevo can be described as a Multi-Method and Multiple Clientele Evaluation Matrix.

Finally, Itai Zak examines methodological problems related to dealing with complex data sets.

The United States' participant, Dr. Robert F. Boruch, of Northwestern University, Evanston, Illinois, makes an attempt to summarize the papers presented by the Israeli authors, and to point out similarities and differences between evaluation patterns in the two educational systems. In his summary article, emphasis is put on communication links established between evaluators and decision-makers in both countries.

It is our hope that this book will be useful to both evaluators and educational policy-makers, and that it will increase the communication between the two parties in a way which will result in the increased infiltration of evaluation results into the framework of decision-making.

Arieh Lewy

ON EVALUATION POLICY IN THE UNITED STATES AND ISRAEL¹

Robert F. Boruch

1. INTRODUCTION

Attempting to understand which of several policies or programs has the greater benefit is not, of course, a novel human enterprise. Comparative tests to understand conditions under which children learn speech, for instance, were undertaken by the Arab conqueror Akbar the Great in 14th century India. Competing theories of human development fired rabbinic argument and theories of evidence during the same period in the Middle East. Nor are sophisticated logic and statistical theories underlying the fair comparison of programs especially new. They are represented in the 18th century scholars' attempts in Europe to understand numerical evidence and independent contributions during the same period to characterize the toxicity of metals, chemicals, and drugs. Finally, there are some distinctive early precedents for controlled field tests of social programs. They include experiments on the effects of sanitation instruction in Syria during 1931-33, and on the comparative benefits of raw and pasteurized milk in nutrition programs for English school children in 1930².

What is relatively novel about evaluation is the regularity of formal government interest in understanding the comparative effects of new social programs and increased government willingness to estimate effects in pilot tests of the programs. This interest in effectiveness is linked in principle to systematically establishing the need for programs and the quality of their implementation. The latter are no less important than estimating effects, but, until recently, had not been routinely required by law.

This paper has two aims, each bearing on recent developments in evaluation policy. The first is to summarize a report that we presented in 1980 to the U.S. Congress and Department of Education, concerning evaluation policy and practices at the national, state and local levels of government. The focus here is on recommendations and the treatment is very brief. The report itself (Boruch and Cordray, 1980) provides details, is readily accessible, and the literature review on which it is based has been published elsewhere (Boruch and Wortman, 1979).

The second aim is to link some of the U.S. recommendations to ideas presented at the Israel-U.S. seminar on education evaluations, and in related papers by the seminar participants. This examination, too, is brief, rather too brief to do real justice to the ideas proposed. But the intention is simply to outline similarities, differences, and analogs between the two perspectives. One of my motives is pragmatic: to learn how the Israeli experience can be adapted by the United States, and vice versa. The second motive is based on the simple premise, suggested earlier, that problems of evidence in this arena are not confined by national borders, ethnic origin, or history. Understanding how durable problems and their solutions are is a task that, as a theoretician, I can ill afford to ignore.

2. THE HOLTZMAN REPORT TO THE CONGRESS

The Education Amendments of 1978 required that the Secretary of the U.S. Department of Education conduct a comprehensive review of federal evaluation practices and procedures. Introduced as a bill by Congresswoman Elizabeth Holtzman, the law directs attention to federally supported programs at the national, state, and local levels of government. In response, two projects were initiated by Department and Congressional staff. A group at Northwestern University was asked to undertake the first in September 1980. The National Academy of Sciences' Committee on Program Evaluation was asked to initiate parallel, independent work. The results of each are reported in Boruch and Cordray (1980) and Raizen and Rossi (1981) respectively³.

The questions covered in Northwestern's report to the Congress and the Department of Education are fundamental. They were implied by the law and the conference reports preceding it:

- . Why and how are evaluations carried out?
- . What are the capabilities of those who carry out evaluations?
- . How are the results of evaluation used?
- . What recommendations can be made to improve procedure or practice?

The study was prospective in its orientation, designed to provide evidence and argument bearing on these questions and to provide recommendations which would help to ameliorate the problems that were identified. The findings and recommendations stemmed from two general sources of information: contemporary investigations by other researchers and agencies, and direct investigations by Project staff. The latter included site visits to local and state education agencies and telephone surveys of local units, both based on a stratified random sample. Round-table discussions were undertaken to

capitalize on special expertise in topics such as school board interest in evaluation. Interviews with some staffers of all major federal agencies with an interest in educational evaluation were carried out. This included the U.S. General Accounting Office and the Congressional Budget Office as well as the education agencies. The literature review covered both unpublished and published documents, including reports maintained by ERIC and, in the case of statute, by the LEXIS system. An earlier review served as a guide to sources on national studies published before 1979.

The project report made recommendations to the Congress and to the U.S. Department of Education. The two sets are condensed in the following treatment and coupled to a brief rationale for each. The links to Israeli work are discussed after each recommendation.

3. REQUESTING AND PLANNING EVALUATIONS

Three of the Report's recommendations concern the process of deciding what kind of evaluations can or should be done, and the way they should be done. They stress the necessity for regular meetings to establish information needs, the merit of specificity in evaluation law, and the reduction of constraints on exchange of information.

On clarifying needs, audiences and options

We recommended that the Congress direct the relevant staff of Congressional committees and support units such as the U.S. General Accounting Office and the Congressional Budget Office to meet with evaluation staff of the Department of Education regularly with instructions to: 1) identify specific committees and groups as audiences for evaluation results, 2) reach agreement about when particular evaluations are warranted and the extent to which each evaluation required by law is possible, 3) clarify Congressional information needs, the quality and type of evidence required, and a planning cycle for each major evaluation required by law, and 4) identify the changes in programs or understanding which could occur on the basis of alternative findings. Parallel suggestions were also made to the Department of Education.

The recommendation is, at its simplest, embarrassingly mundane: it asks that the principals meet. And indeed periodic efforts have been made by Congressional and Department staff to assure that the production of reports coincides with authorization cycles and that Congressional needs are understood. But the process has been less orderly, less regular, and less thorough

than it ought to be. This recommendation depends heavily on the fact that a legislative demand for "evaluation" is often ambiguous. The word can imply any activity, from journalistic reporting to full-blown long-term field experiments dedicated to estimating the effects of innovation on children. The involvement of multiple interest groups is often necessary. But this complicates matter further since all are unlikely to agree on just what sort of evaluation is warranted. At worst, general legal demands to evaluate that are unaccompanied by serious discussion obscure the fact that the feasibility of particular kinds of evaluation varies enormously and that elaborate evaluation may be unwarranted.

Specification in law

We recommended that in constructing laws for evaluation the Congress: 1) specify exactly which questions ought to be addressed and the audiences to whom results should be addressed, when specification is possible; 2) provide for formal assessment of the evaluability of the relevant program where specification of questions is not possible; 3) provide for statistically valid field testing of proposed evaluation requirements where specification is not possible and in-house assessment insufficient.

Though statutes are frequently explicit about routine reporting requirements, references to evaluation are often ambiguous. The common requirement, for instance, to evaluate "whether the program meets the objectives of the statute" is common but vague. The published Hearings, covering public testimony submitted prior to enactment of a law, are not always informative.

Defining evaluation requirements in terms of the questions that should be addressed is sensible so long as the questions themselves are clear, answering them is feasible, and the answers are likely to be useful. The particular questions that often need to be addressed are: How many are served and how many need service? What are the services and their costs? What are the effects of programs on their primary or secondary clients? What are the costs and benefits of services? The early specification of audiences, especially particular committees or Congressional support agencies, should enhance the usefulness of reports.

We recognized that explicitness in law is often not feasible or desirable. Consequently, we suggested formal investigation of evaluability (Wholey, 1977) to clarify questions, audiences, and the ways in which results could be used, within a year after enactment of a demand for evaluation. We recommended field

tests of reporting requirements in the interest of assuring that costs and benefits of reports, users and uses of reports were well understood.

Authority for technical discussion

The third recommendation in this class urged that the Department authorize the technical staff of evaluation units to initiate discussion of evaluation plans with pertinent Congressional staff, at their discretion, and to refrain from directives which might impede direct discussion.

The impetus for the recommendation was simple: Competent evaluators can expect to do a good job only when they have the opportunity to frequently discuss Congress's information needs. Restrictions on the evaluation unit's initiating discussion with the Congressional staff of committees that demand evaluation prevent the job from being done better. Such restrictions were made formal by, among others, Joseph Califano during his tenure as Secretary of the Department of Health, Education and Welfare. The Report recognized that some restrictions on bureaucratic lobbying for programs are warranted, and that some administrative rules are necessary to keep the process of communication between agencies and the Congress orderly. Restrictions engender a lack of clear opportunity to identify which information Congress can use. This in turn decreases the likelihood that evaluations will be timely, relevant, and credible and the likelihood that the Congress will find the results useful. Relaxing restrictions will not, of course, guarantee usefulness.

Remarks

Three aspects of the Israeli papers are pertinent to these recommendations.

Mapping the Question. The first concerns Louis Guttman's mapping sentence which, as Lewy describes it, is a remarkably terse statement for helping one identify critical decision points in the evaluation: when information is warranted (at what stage of the program), what entity ought to be evaluated, why the information is needed, how it should be obtained. This literal map is implicit in our first two recommendations to the Congress. Moreover, it constitutes a neat working rule for the individuals responsible for planning evaluations at the national level. It takes no wit to see that it can be formally adopted as well in work at state and local levels of government.

Audiences for Results. The mapping sentence dedicates no explicit attention to the matter of whose needs must be served by the evaluator, but both the Holtzman Report and the papers in this volume do so. For instance, one

of our major suggestions was to identify audiences for results as soon as possible. At its worst, this is merely pious exhortation: the turnover of staff members of the Congress and at the executive level is high enough to threaten some short-run projects and most long-term research. But it is a practical suggestion to the extent that career bureaucrats and Congressional committee staff that are responsible for evaluation are a stabilizing influence and can serve as a vehicle for identifying both transient and durable users of information.

The idea of regular meetings among both evaluators and users of information is not different in principle from the tactic already used at the Israel Curriculum Center, judging from Lewy's paper. The ICC's use of a liaison person as a bridge to users and as a expeditor seems sensible for information exchange, building trust and a common vernacular. Note that our recommendations, though, concern only evaluators and users. Production or development agencies are ignored. In principle at least, the ICC liaison approach is adaptable to working relations between these two groups in the U.S. as well at the federal, state, and local levels.

Developing an Evaluation Portfolio. Lewy cites a 1970 article by Alkin, suggesting that the task of evaluators is in no small measure, "ascertaining the decision areas of concern". Lewy expands on this to argue that the evaluation unit adopt this as a fundamental operating principle, and moreover that the selection of the topic for evaluation "should be done on the basis of consent between or at least compromised among the two teams", the two teams being the evaluation unit and the program development team. The Kugelmass discussion of the Reform-Junior High School change instituted by the Israeli Knesset makes a similar point. In 1968, the law altered the then conventional 8 year primary school and 4 year high school program into a new 6-3-3 sequence in government schools. Kugelmass suggests that there was a great deal of difficulty in meetings among the administration of the Ministry of Education and researchers, and stresses that making a decision about what sort of evaluation to do, under differing pressures from diverse interest groups, is not simple or easy. Dan Davis too makes the point. But he emphasizes that once a preference is made explicit about the desirability outcome evaluation at least, the evaluation must be under considerable control by the evaluator to assure a reasonably successful evaluation.

The problem here is not new, of course. It underlies any attempt to build a coherent research and development agenda, any effort to choose among products for manufacture or corporations for acquisition. Nick Smith's labelling of

the process of trying to choose among evaluation enterprises as portfolio development is also apt. (The label may assist in translating ideas about evaluation to a nontechnical audience, such as a legislature, that is sometimes more sensitive to business than to the evidential basis for government).

Nor is the difficulty of choosing what sort of evaluation to perform confined to Israeli borders. The U.S. encounter with the same problem is one of the reasons for making clear the choice and the basis for choice in the Holtzman Report's recommendations. The pertinent evidence comes from a variety of sources. Mary Kennedy (1980) for instance suggests that efforts to compare the relative effectiveness of two or more strategies of (say) instruction are not common at the school district level. Her message is that impact assessment is less frequent and less important than other evaluation questions in the local agencies. Charles Stalford (1980) quite properly warns against a "testing-only model of evaluation" (p. 6). The Holtzman Project and other work seems to support this contention: activities other than impact estimation are important and the importance varies with the level of government and with the agency within government.

This of course does not mean that comparisons are unimportant, merely that they can receive low priority. The reasons may include simple inability to create variations that are cheaper and more productive than the existing one and that are worth testing. For instance, in response to my suggestion that the Agency for International Development test variations, one AID staffer complained plaintively that they had had enough trouble creating one variation and that creating more just to be able to find the most effective one was too onerous to countenance.

The more general implication is that within a school district or at any other level of government a de facto portfolio of evaluation activities is created. The question this engenders is how such a portfolio can or should be built. Consider for instance the first factor that might be taken into account when developing a portfolio: the source of the inquiry or target audience for evaluation results. Is it sufficient to rely solely on evaluative questions from instructors, parents, or program managers to develop a portfolio? Probably not, since asking the right questions requires some skill and informed conceptions of evidence. Can one rely solely on the evaluator? Probably not, since this assumes too much knowledge of substantive problems. What do the stereotypical portfolios look like in this respect? The U.S. General Accounting Office initiates about two-thirds of its own inquiries and most of these are managerial studies. Should evaluation offices

at the state level build a portfolio in the same way? We know very little about this. Nor can we give much advice.

The second factor is time. We know that fast turnaround studies are essential to satisfy a public or a superior with a short attention span, if not to actually resolve durable problems. And so perhaps most evaluative studies need to be short in the interest of evaluator survival. But all evaluators, especially those in government and academe, do have some responsibility for finding long-term effects of programs, and for understanding long-term social problems as well. There is no technology for designing evaluations which produce short, interim, and long-term results.

The strategy has been to elicit suggestions for evaluations from directors of substantive programs. The ultimate choice is based partly on agreement between evaluators and these agencies in principle. But it may be superseded by agreements with the Secretary of Education or by other criteria used in making decisions. The other criteria include expiration dates for legislation bearing on programs, the period during which a legislative committee could be expected to use information or whether high priority programs have been evaluated. The choices are incorporated into three plans for evaluation.

Very little intellectual attention has been dedicated to a third factor, the administrative mechanism which yields the evaluation portfolio, and which can be used to terminate projects which are not turning out well. The system at the U.S. General Accounting Office appears to be hierarchical, involving screening committees to ultimately determine the choice of project, and a special committee for termination. Nomination of topics to be investigated come from staff groups with operating responsibility. Until recently, the Office of Education and Dissemination at the Office of Education had a similar system. But it is not clear how termination decisions were made.

The main point is that criteria for developing and assessing the value of an evaluation portfolio are not yet clear. Serious attention has been given the matter by bureaucrats, not by executives or academics, and their effort can be augmented profitably by others.

4. EVALUATOR CAPABILITIES

The Project staff had been asked to investigate the capabilities of those who do evaluations and to make recommendations based on our findings. We recommended that the Congress and the Department: 1) assess capabilities of local and state education staff before new statutory evaluation requirements

are directed at them in order to determine where resources are adequate to meet the demand, 2) expand training or technical assistance when the demands are notable and capabilities low, and 3) explore the feasibility and desirability of direct contract programs to capitalize on capabilities in strong local and state education agencies.

The first section of the recommendation stems partly from the fact that no real standard for assigning the title "evaluator" exists and that skills required of the evaluator depend heavily on the nature of the evaluation demand and on local and state interest in evaluation. The second part is based on the finding that most local and state agencies need assistance when the evaluation requirements are technical. The minority of these agencies that do have strong evaluation units are a major resource, and we believe that direct grant opportunities should be expanded to capitalize on them.

By determining capabilities here we mean understanding whether there can be a reasonable match between what the law demands of local and state evaluators and the skills of these individuals. A formal assessment of this sort is unlikely to be easy for three reasons. First, within a school district or state office, evaluation responsibility may be split up among several individuals, none of whom may have any pertinent training, and this responsibility can often change. Second, evaluation duties may have a considerable range depending on local interest in exploiting systematic information to improve programs. Just meeting minimum federal requirements requires far different resources than establishing a long-term research program. Finally, the methods one might exploit to perform capabilities assessments are not clear. They range from intensive task analyses during, say, pilot tests of new regulations that require a specific type of evaluation to telephone surveys that enumerate skills and tasks. Drs. Georgine Pion and David Cordray are developing plans now to accommodate such problems and to implement assessments for local education agencies and for community mental health centers.

A critical influence on the matter is whether an education agency decides to just accommodate federal requirements or goes beyond these to mount a stronger evaluation program. Even just meeting requirements demands some skill. The notion of temporal instability or reliability of tests is not obvious to many people despite training in a substantive education area and in the history of testing. As a consequence, we urged that the federal program sponsor make an effort to capitalize on evaluation unit expertise in training local or state staff in meeting demands, and that funds be allocated

for such training. Options such as cooperative arrangements among small education agencies for joint support of an evaluation unit and expansion of federally supported technical assistance centers need to be explored, and the Report suggests doing so.

For local and state agencies that are willing to go beyond federal requirements, we stressed direct grants from the federal government for two purposes. First, some agencies are capable of mounting research and evaluation programs that match federal efforts in quality and are more pertinent to local interests. They are in the minority, accounting for probably no more than 400 of the 15,000 school districts and less than half the state agencies, and they deserve to be given an opportunity to produce good work that can be applied to other areas. The second purpose is to foster closer ties between local agencies and university evaluation groups. Such arrangements are bound to be difficult, but it is hard to see how the state of the art in evaluation can be advanced without better ties between the two.

Remarks

There are several points of correspondence between the findings on which these recommendations are based and the opinions registered in the seminar papers. Consider, for example, Lewy's conclusion that "a realistic assessment of actual need in terms of manpower and other resources and their satisfactory provision constitute a prerequisite for the successful operation of the evaluation unit. It is not the absolute size of the budget which determines the successful operation... but rather the match between the resources available and appropriate definition of the evaluation tasks". This is remarkably similar to the conclusions tendered by Pion and others in the Holtzman Report that evaluation tasks vary widely among school districts, that the skills required to do those tasks vary as well, and that the tasks have to be understood before resources can be intelligently allocated to training and before laws demanding wholesale evaluation can be conscientiously constructed.

The second point of correspondence lies in Kugelmass' observation that as a result of the academic emphasis on basic research and theory, the manpower available for task-oriented research such as evaluation is spare. Tamir points out that even where university-trained researchers are available, there will be a notable tension between the research-oriented view of what should be done at what level of accuracy, and what the manager or practitioner believes

is warranted. That same problem has appeared in the United States and is being resolved in several ways. A tight market in university jobs seems to have resulted in better people going into government, into independent contracting research institutes, and into evaluation units at the local and state levels. The migration engenders problems but it is also reasonable to expect better understanding of local practical problems and wiser evaluators. The federal government has assisted by creating technical assistance centers to respond solely to local needs for advice. The centers are staffed by university trained people; not all local programs, however, are assisted by such centers.

A third point of correspondence is Lewy's conclusion that the evaluation unit serves as a "catalyzer for initiating evaluation activities, the limits of which exceed the working capacity of the unit itself". This suggests that developers get interested in evaluation to the extent that the working relationship with evaluators is close and that this interest can be used to expand the effective size of the evaluation unit's staff. There is an analog here to efforts at the local and state levels in the U.S. to augment the evaluation requirements set out by the federal government. In particular, though only a minority of the evaluation units and research units within school districts are strong, this minority used the minimal requirements as a vehicle for collecting additional information of more direct relevance to local interests. States such as California and Massachusetts also have this utilitarian perspective, building on federal investments and requirements to do a better job in meeting federal, state, and local demands.

One feature of the Israeli experience for which there is no routine analog is the use of a liaison person to link the program development group with an evaluation group. Informal arrangements of the sort do appear in the U.S., but the role seems much better articulated in the ICC. Lewy's description of the liaison person's training and skill is especially interesting. If I understand it correctly, the three types include those with evaluation training of a substantial sort, those with substantial substantive training, and the project director. Lewy suggests that the project director does not work out too well because he hasn't got the time, and that the substantive area expert is probably best because he is immersed in the project itself.

5. DESIGN AND EXECUTION OF OUTCOME EVALUATIONS

Once said, it is obvious that quality in design of an outcome evaluation affects quality of the data and of conclusions. The evidence that bad design

can make programs look worse than they are, or better than they are, or yield ambiguous evidence is substantial. The theory that organizes understanding of biases in estimating program effects is, however, reasonably well articulated.

The idea that quality in design ought to be recognized as a formal part of evaluation policy is explicit in federal education agency attempts to yoke the introduction of new programs with design, as in evaluation of the programs supported under the Emergency School Assistance Act, in attempts to review designs with more vigor within agencies, and in efforts to provide technical assistance programs in the interest of better local design. It appears also in the U.S. General Accounting Office's attention to competing explanations characteristic of poor designs, to the elements of reasonable design, and to the need for designing evaluations before a new program is put into the field. It has been recognized by the courts in cases outside education which recognize the flaws in some evaluation designs and the benefits of others. The Supreme Court's Federal Judicial Center, for example, is developing policy on the use of randomized field experiments in legal settings to make clear the issues and precedents. The theme of quality, though, is not sufficiently well established to flourish without periodic reiteration. The task was undertaken in the Holtzman Report through two recommendations, one made to the Congress and one to the Department of Education. A third, concerning standards, is treated later.

Pilot tests and designs

We recommended that the Congress: 1) routinely consider pilot testing every major new program, major variations on existing programs, and major program components before they are adopted at the national level, using high quality evaluation designs, and 2) authorize the Secretary explicitly, in each statute that requires estimates of the program's effects on target individuals, to use high quality designs, especially randomized field experiments, for planning and evaluating new program components, program variations, and new programs.

The rationale for the first part of this recommendation is that higher quality evaluations are more feasible before the program is adopted at the national level. Political-institutional constraints are likely to be less severe, better designs can be employed, and conclusions then are likely to be less ambiguous. The introduction of new programs can be staged so that earlier stages are pilot tests for later ones. We stress formal tests of

new program components and new variations here because such evaluations are not a matter of common practice.

The second part of the recommendation, as well as the first, stems from our conclusion that better designs must be used if the Congress or the Department wants good estimates of the effects of programs on children. We do not advocate estimating those effects in all cases. Estimation is complicated under the best of conditions, despite simplistic announcements that the "program was successful because test scores went up" or that it was unsuccessful because they went down. Nor do we believe that designs that are high quality relative to statistical standards are always feasible or warranted for estimating program effects. We do advocate explicit authority in statutes for high quality designs, especially randomized experiments, to facilitate their use. We believe explicit statutory provision is essential because such designs are the best in principle, and that should be recognized. The authorization should provide for review of the use of these designs.

Tests of new program components, program variations, and new programs

We recommended that the Department of Education explicitly authorize the use of high quality evaluation designs, especially randomized experiments, in evaluating new program components, new program variations, and new programs, in all regulations that require estimating the effects of innovative changes.

The main justification is that high quality designs lead to less debatable estimates of programs on children than do low quality designs. They are less difficult to execute and are more feasible for pilot testing new programs, program variations, and program components, than for estimating the effects of ongoing programs. Explicit authorization would make the importance of good designs plain, and would provide a more clear opportunity for competent state education authorities (SEAs) and local education authorities (LEAs) to exploit them. We use the word "authorize" here rather than "require" to make clear that the evaluator is empowered to use an experimental design but need not do so if it is unwarranted or not feasible.

Remarks

The recommendations on randomized field tests were supported by some evidence on their feasibility and appropriateness. A judgement about feasibility in the particular case, we believe, should be based on precedent, for a number of field experiments have been undertaken in education and other

areas. It should be based, for complex evaluations, on pilot tests of the experimental procedure itself, for one cannot anticipate all problems engendered by the method. And it should be based on independent criteria such as whether the service is in short supply and randomized assignment is indeed an equitable method of allocating it. These criteria need to be explicated better, and they need to be linked to broader testing strategy.

Standardized Evaluation. In doing both, we can rely partly on Davis' presentation. He proposes that six conditions must be met in order to obtain decent estimates of the effects of programs. 1) The program has to have relatively clear goals and operating procedures, that is, it must be implementable. 2) The evaluator must be responsible for both the program operation and its evaluation, maintaining special control over evaluation. 3) The program must be implemented first in an optimal setting - field conditions, training, and the like being the best possible. 4) Schools must be selected for their willingness to participate in the research. 5) The research design must approximate laboratory models in terms of assignment and execution of evaluation. 6) The results of the evaluation must serve as a standard against which normal field operations can be judged.

Not content to just lay out conditions, Davis is attempting field trials under these conditions on a program that has never been investigated well, despite its attractiveness, in the U.S. or elsewhere. The program is, as I understand, a national tutoring effort in which university students get academic credit for helping children in grades five through nine. The regular program has minimal supervision and the more elaborate "optimal" version involves intensive supervision and more hours of tutoring. The more intensive version is distinctive in that "the program is a supervisory and guidance structure which is sensitive to the problems encountered by the tutors and can help in solving them, and the program is maximally flexible so that it can adapt to the specific conditions of each tutor-tutee relationship". Earlier evaluations of the ongoing program show mixed results. Consequently, a good field test of the ideal version of the program is a natural way to understand what the maximum effects of tutoring can be.

Apart from the conditions that Davis proposes, his general strategy of conducting a controlled experiment of an optimal program to gauge the maximum effects of subsequent or ongoing programs is an attractive one. It is generalizable to the U.S., at least in principle, and does not appear to have been suggested before, at least not as explicitly. There have been related suggestions however. For instance, the Riecken et al. (1974) volume on

social experimentation stressed the idea that in field tests of programs, one ought to assess not only program levels that are clearly feasible, but also some that are not practical at the national level. The argument is based on the premise that "practical" programs are often weaker than we expect them to be and that high dosage (optimal) programs, though impractical at one time, may be practical in the future. This is especially likely if one finds that the higher intensity does produce notable effects while low dosage "practical" programs have no detectable effect at all.

Testing Components and Variations. No theory of evaluation demands that the effects of an entire program be estimated, and few practitioners would regard such an unqualified demand as sensible. Yet professional vernacular, rhetoric, and legal mandates foster the view that wholesale evaluation is warranted, distracting attention from the possibility of testing components of programs. For example, one may find that running high quality tests of new parent education programs is not possible. But estimating the effect of alternative sources of information, of various ways to present the information or ways to prevent ingenuous use of information, and so on, may be possible in small high quality experiments. The strategy of component-wise evaluation has been exploited in the U.S. evaluations of the Emergency School Assistance Act, in research which preceded the development of Sesame Street, and elsewhere. Incorporated into evaluation policy, the idea broadens early options, and in the event of a major evaluation's failure, it is a device for assuring that at least pieces of the program can be assayed properly.

Analogous to this approach are not difficult to find in work at the Israel Curriculum Center, the High School Biology Project, and other projects. The process of identifying which components to evaluate seems to have been routinized best at the ICC, notably by exploiting the mapping sentence approach: Lewy's application suggests that focusing on the entity to be evaluated is integral to the continuous evaluation strategy. The Israeli adoption of high school biology curriculum (BSCS) programs developed in the U.S. is pertinent too. According to Tamir, the BSCS program was available in three versions. The blue version emphasized bio-chemical concepts and is relatively sophisticated. The green emphasized an ecological perspective. The yellow stressed a more conventional approach, but was somewhat more interesting and adaptable to the Israeli perspective. The important point here is that three variations of the program were developed. This seems an imminently sensible idea when there are major differences in perspective

about how something should be taught and major differences in teacher opinion about how it can be taught. Moreover, the variations' relative effectiveness are testable in principle. In practice, pieces of each are likely to be evaluable.

Dan Davis's "standardized evaluation" is also consistent with this theme. To determine the maximum possible effect of an ongoing program which itself may be difficult to evaluate well, one may invent a very intensive variation on it, an optimal version, and submit this to very well-controlled tests. The idea can help to circumvent the chronic problems of estimating the effects of ongoing studies.

Finally, the plan being developed by Gershon Ben-Shakhar and Baruch Nevo for understanding the effectiveness of matriculation tests reflects some of the same spirit. Formal testing in a part of a large, complex education system and its evaluation, more or less independently of the rest of the system, is an idea worth exploring. The so-called Irish Study has had a distinctive advantage in this regard, since standardized testing is not common in Ireland and evaluators could introduce it on a trial basis and estimate its effects on teachers and students using randomized experiments and other evaluation designs (Airasian, et al., 1978).

6. CRITIQUE AND SECONDARY ANALYSIS OF EVALUATION RESULTS

The evaluation design, its execution, and the skills of original investigators are basic to the production of useful information. But they are not always sufficient. The pressure toward less than candid reporting is sometimes great, and it is not always clear that one can resist them. Egregious errors are made and corroborative or contrary evidence is ignored, for the time available for analysis is not always adequate. The benign skepticism necessary for the in-house evaluation generates a reasonable but parochial picture. A less benign, or at least more impartial, outside analyst could come up with different conclusions. Finally, data generated in social program evaluations constitutes a national resource and should be treated as such. The research can be expensive despite the production of data that are useful in short-term decisions. It behooves the evaluator to learn how to exploit the information repeatedly.

Partly for these reasons, we recommended that in statutory requirements for evaluation of major programs, the Congress: 1) require an independent, balanced, and competent critique of evaluation results that are material

to policy decisions, 2) require critique of samples of evaluations submitted by LEAs and SEAs in response to legal requirements, and 3) require that statistical data produced by national evaluations be made available for reanalysis.

A complementary recommendation was made to the Department of Education: 1) Incorporate into procurement procedures and policy the requirement that all statistical data produced in major program evaluations be documented and stored for reanalysis. 2) Create an administrative mechanism for deciding when simultaneous analysis by both the original evaluator and an independent analyst is desirable and feasible, and a mechanism for executing simultaneous independent analyses.

The text of the Report made it plain that we did not mean adverse commentary in using the word "critique". The idea is to ask for reasoned judgements about whether conclusions drawn from the evaluation are sensible and can inform decisions. The immediate reason for the recommendation is that such criticism is not routine but is essential to enhance the credibility of good evaluations, to properly identify poor evaluations as such and to provide feedback to federal evaluation units, contractors, and grantees about the quality of their work. There is no formal system for the competent critique of evaluation reports produced by local and state education agencies in response to federal law, yet many such reports could benefit from conscientious review.

The elements of a system for critique and secondary analysis should include: 1) an explicit institutional policy on the rapid disclosure of reports and access to the data underlying the reports, 2) a mechanism for independent critique or secondary analysis where possible during an evaluation, and where this is not possible, after a report is submitted formally, and 3) guidelines on the reporting and storage of information.

The Report recognizes the problem that criticism may be witless and counter-productive. The recommendation is based on the premise that the long-run benefits will offset the effects of self interested criticism and the burden that criticism imposes on the evaluator.

Remarks

No explicit attention to this matter is evident in the seminar papers. Rather, the theme is implicit in the spirited exchanges of opinion during the seminar, and in custom if I judge correctly the participants' stress

on presenting research results in the sometimes harsh climate of professional forums. The High School Biology Project, for instance, appears to have generated a large number of fascinating papers that are informative to researchers inside and outside Israel. As I understand it, the Office of the Chief Scientist too can serve as a device for independent critique, and perhaps secondary analysis as well. But the ordinary mission of the office, in advising and deciding on research projects, can dilute administrative independence. Similar offices in the United States, such as the Director of the research-oriented National Institute of Education, operate with a similar constraint: fiscal, administrative, and bureaucratic independence is a matter of degree.

Reasonable critique depends on standards, and standards were addressed in both the Holtzman Report and in the seminar papers. They are considered in the next section.

Analysis and Competing Models. The Holtzman Report did not examine the methods of data analysis used in evaluation. Its principle audience was nontechnical and most of the questions it addressed were answered in a non-technical way. Other research produced by North western, however, has examined technical issues in assuring access to and the quality of data for analysis and the nature of competing analyses that might be undertaken. The volume edited by Boruch, Wortman, and Cordray (1981), for instance, considers both policy and practice, and can be regarded as an explication of the Holtzman Report's recommendations on secondary analysis.

Itai Zak's paper is most pertinent to this level of detail. It represents a statistical tradition of trying to understand the structure underlying data, of establishing the extent to which a theory represented mathematically is consonant with the information. The tradition is represented most visibly in econometrics, but recent attempts by Goldberger, Jöreskog and others to link approaches in psychometrics and econometrics through structural equation models, and the quantitative sociologists' work on the latter have led to some remarkable advances in facilitating and understanding their use. Several points bearing on the general topic seem worth making. They are implicit in Zak's approach.

First, methods such as Wold's PLS and Jöreskog's maximum likelihood approaches permit one to relax the assumptions characteristic of conventional textbook approaches such as regression analysis. Simply put, they allow one to build more realistic models of reality. That this advantage is not trivial is apparent from policy-relevant research in the U.S. For example, the

Westinghouse-Ohio State University evaluations of Head Start, a preschool program for deprived children, resulted in estimates of program effect that were near zero and in some cases negative. At worst, it seemed that the program hurt rather than helped in their conventional covariance analyses. Secondary analyses of the same data, by Magidson (1977), capitalized on structural equation models similar in character to those that Zak uses. The new results, based on less demanding assumptions than the original work, suggest that the program had positive though small effects on the cognitive ability of children who participated.

The second point is that this benefit of new methods also produces ambiguity. Disparate models, theories of behavior if you will, may fit the data equally well. So, for instance, the Magidson results are being debated by other analysts who believe their models are at least as appropriate. This ambiguity is typical. And consequently it behooves the analyst to fit several models to his data, in much the same way that Zak has done.

Another point worth recognizing stems from the fact that these new model-fitting approaches have developed independently of methods in conventional randomized experiments. The two cultures here are different, but failing to recognize their linkages would be a mistake. It is possible, for instance, to express the ordinary analysis of the variance model that underlies randomized tests in terms of structural models. Lee Wolins and I have done so for one class of models but not much other work seems to have been done. Perhaps more important, the structural models lend themselves to internal analyses in experiments. That is, having discovered that a program has an overall effect, based on randomized design and conventional analysis, one may exploit the new methods in path analysis to better understand links between specific components of the program, specific types of participants, and specific outcome variables. Something of the sort has been tried in an analysis of prison parole programs by Rossi, Berk, and Lenihan (1980). More distantly related approaches are not uncommon in analyzing the results of field experiments on income subsidy programs for the poor (see Boruch, Cordray, and Wortman (1981) for other illustrations).

7. STANDARDS AND GUIDELINES

One of the justifications for the Holtzman Project was Congressional interest in whether evaluations could be subjected to uniform standards for judging their quality. Indeed, a variety of guidelines to judge quality have been developed by the U.S. General Accounting Office, the Evaluation

Research Society, and the Joint Committee on Standards for Program Evaluation, a group whose members include representatives of most professional associations with an interest in evaluation. Crude standards are also embodied in certain federal activity, notably the Joint Dissemination and Review Panel, whose mission is to assess the evidence on locally developed programs in order to determine whether the programs merit federal support for distribution to other local agencies. There is substantial overlap in topical coverage among guidelines. But they differ in detail. Our review led to the following recommendation: while recently developed standards and guidelines for evaluation should not be incorporated into law, they are sufficiently well developed to recommend that the Congress: 1) use such guidelines to understand what can reasonably be expected of evaluations, 2) direct that agencies use them as a guide where appropriate to developing criteria for judging evaluation plans submitted by local and state agencies, and 3) elicit assistance in the interpretation of guidelines from Congressional support agencies, such as GAO, that have been instrumental in their construction.

The main reason for recognizing that guidelines be recognized officially is that we believe they can be useful in explaining what is meant by evaluation to the public and its representatives, and in informing the public about what can reasonably be expected of evaluation projects. Guidelines may also assist in protecting the competent evaluator from incompetent criticism. They should certainly help one to identify inept evaluations.

We argued against incorporating such standards into law because neither evaluation law nor the standards are sufficiently well developed as yet to justify incorporation. Moreover, giving legal status to specific guidelines can impede the development of better guidelines, are almost certain to be applied inflexibly, and are likely to do more damage than good in other respects.

Remarks

Most standards are general and their proper application depends on circumstance. A survey of needs, for instance, requires a subset of criteria in a complete list; a randomized experiment requires a few of the same criteria but others must be employed as well. The Holtzman Report did not address this matter in detail. Instead, it focused on when guidelines should be used, notably in judging the merit of grant proposals and contracts

that contain evaluation plans and, once the program is evaluated, in judging the quality of the evaluative evidence.

The ICC evidently takes a somewhat different approach in ascribing the idea to "minimal evaluation requirements". As I understand Lewy's remarks, the concern is not only on when evidence becomes material but on the general kind of evidence as well. Expert judgement is regarded as essential for judging the quality of materials, observation of the teaching and learning processes are essential during early try-outs of the material in classes, and assessing cognitive achievements of children is essential at the end of the first try-out. The first two activities are the responsibility of the program developers and the last is undertaken primarily by the evaluator.

The ICC criteria are not incompatible with other, more elaborate guidelines. They can be regarded as a distinctive operationalization of items that appear in more general lists and, partly because of their brevity, are likely to be a useful operating rule in at least some large local education agencies in the U.S. And there are distinctive parallels to this minimalist approach in some U.S. evaluation agencies. The Officer of the Inspector General in the Department of Health and Human Resources, for example, undertakes fast turnaround studies that rely heavily on expert bureaucratic judgement and some on-site observation of processes in, for example, health services delivery (Hendricks 1981). But his effort is dedicated to administrative failures rather than to programs in general. There is some similarity also to recent Agency for International Development efforts to execute fast turnaround studies of foreign assistance projects. Here, as in the Israeli case, there is often little time to do much more than obtain expert judgement and crude observations.

Implicit in both the Holtzman Report and in Lewy's standards paper is the idea that guidelines for judging quality can range considerably, from the very permissive to the very general. The clearest illustration of both the idea and its exploitation comes from the engineering sciences. Here, the demand level of the standard depends on the uses to which the information is put. Relatively wide tolerances are permissible in local geophysical measurements since the information is used primarily by lawyers and construction engineers. Much closer tolerances are required in geophysical measures made for some scientific studies, on land erosion for instance. The same spirit is evident in the ICC's minimalist approach - expert judgement normally being less precise and certainly less verifiable than elaborate

observation. It is evident too but underexploited in the U.S. It is sensible to consider using one set of standards in special grants made for small innovative programs and a less restrictive set for programs that operate with regular budgets. It may be sensible to require larger, more capable school districts to provide information that accords with more rigorous standards and to require much less precise information from the less capable ones. The implications have not been worked out nor have the options been articulated well in evaluation management in the U.S. or abroad. Both tasks should be undertaken.

8. THE USE OF EVALUATION RESULTS

Whether the results of evaluation are used and by whom they are used was a fundamental concern of the Project. We uncovered a great deal of information on use, but the research was difficult. Not the least of these was lexical confusion. A federal director of research, for instance, announced that he did not perform evaluation at all despite a list of projects under his direction that included evaluations labelled as such. His superior, interviewed fifteen minutes later, claimed the contrary: that the division produced a great many evaluations all useful to management. We encountered a Congressional staffer who announced baldly at the beginning of an interview that his committee did not use evaluations, yet he later said that evaluation reports were used to guide committee hearings on programs.

This confusion, or at least inconsistency, underlay a good deal of debate about the utility of evaluation results. And so we defined "use" explicitly to mean: 1) applying these results in making specific decisions about law, regulations, budgets, or related administrative topics, and changes in substantive content of programs, 2) capitalizing on them to enhance understanding of issues even where a decision could not be made, or 3) exploiting the information to persuade others, as in political speeches, or to confirm one's own beliefs. The Project's efforts to document use and nonuse of evaluation reports focused on specific evaluations and stressed the corroboration of evidence from different sources. The findings suggested, as one might expect, that some evaluations are used and some are not, and that use depends heavily on the planning of use, close relations between the user and evaluator, and willingness and capacity to use results.

Our first recommendation bearing on use of evaluation results was made to the Congress. We urged that its members:

- 1) direct the staff of relevant committees, the Department, and the GAO to routinely outline which institutions can reasonably be expected to use results of each major evaluation and how such results might be used, during the design stage of every major program evaluation;
- 2) specify exactly which evaluations have been used and why they were used, which have not been used and why they were not used, in authorizations and appropriations committee reports;
- 3) require evidence about specific changes resulting from evaluation, whenever the law requires state agencies to describe uses of evaluation; and
- 4) explore the feasibility of direct competitive grants and contract programs focused on improving the use of results at the local and state education agency levels.

The origins of the first part of the recommendation lie in the absence of any mechanism for planning use at the national level. Simply put, unless specific user groups are identified and some decision options laid out, evaluation results are less likely to be used. Indeed, if there is no clear way to link the evaluation with decisions or considerably better understanding, one can argue that the evaluation should not be performed at all. Specifying users and options will also help to make it easier to track utilization, and that, in turn, will help to inform judgements about how evaluation resources could be better allocated. The recommendation to cite useful and useless evaluations in federal reports and to require SEAs and LEAs to record specific changes has the same objectives: better understanding of use and better resource allocation. The suggestion to identify useless evaluation is not an invitation to criticize arbitrarily. We found that some local and state education agencies are capable and interested in inventing and testing better ways to use information. The suggestion to expand their opportunities for doing so is based on this.

A second, related recommendation was made to the Department of Education. The Report urged that evaluation unit staff and evaluation contractors be directed to 1) provide oral reports regularly as well as written reports on results of major evaluations, and on the uses to which results can be put, to relevant Congressional staff and support agency staff and the program staff within the Department; 2) create a system to periodically collect, synthesize, and report specific uses to which evaluation is put; 3) improve the Annual Evaluation Report by citing instances of use more specifically;

and 4) direct evaluation staff to meet regularly with Congressional staff to clarify information needs, feasibility of evaluation, audiences for results, and ways in which results can be used to modify programs.

This cluster of suggestions is based partly on the finding that the use of evaluation results is not tracked conscientiously and the belief that it ought to be tracked to learn how to perform evaluations better, and how to better allocate evaluation resources. The rationale for the last recommendation is identical to the one given earlier for the Congress on planning and executing evaluations.

The final suggestion under this topic focuses attention on assuring access to and better specification of reports.

We recommended that the Department 1) adhere to a clearance rule which makes evaluation reports automatically available after a fixed number of weeks; 2) specify completely the evaluation documents referred to in the Department's Annual Evaluation Report, the Federal Register, and policy statements; and 3) include, in every major evaluation report, a list of core recipients.

The recommendation stems partly from difficulties encountered in obtaining reports under review by the Executive Secretariat of the Department of Health, Education, and Welfare and other groups involved in the DHEW clearance process. We also found it difficult to identify reports precisely when they were cited as evidence of the usefulness of evaluation in developing regulations or policy. The absence of a list of core recipients of reports made it very difficult to identify potential user groups and to determine if reports were used. The consequence is that what is useless or useful is less verifiable.

Remark

The basic idea that there needs to be a group constituted to reason from the data is implicit in both our recommendations and in Israeli operations. The form differs a bit, though.

Committees to Reason from Data. Consider, for instance, Kugelmass' description of the Van Leer study of primary schools, a massive undertaking to understand the process and products of primary schools in Israel, including controversial issues such as integration and religious vs. non-religious school systems. One distinctive aspect of this enterprise was that the Ministry of Education and Culture and the Chief Scientist took pains to

appoint a high level committee to understand the ways in which the results of this research could be used by the Ministry. The Chief Scientist, as well as heads of various divisions of the Ministry and of the primary school system, were involved in the committee, and a number of sub-committees were set up to recognize special interests and problems. A mechanism of this sort was not suggested in the Holtzman Report. But it does seem to be a sensible way to understand how policy implications can be deduced from such data. The Holtzman Report assumed that negotiation and regular communications between the Congress and the Department would facilitate in understanding the uses to which the data could be put. The power that a blue ribbon committee has to do this, or to exploit suggestions made at lower levels could be used to make negotiation more effective.

Decision Options. The Holtzman Report marshalled a good deal of evidence on the use of evaluations in decisions at the federal level. But we obtained very little for local levels because of the difficulty of corroborating use. Nor did we classify the uses according to what decisions they might concern. Partly for this reason, the classification schemes developed by Lewy and Davis are pertinent. They are tidy ways to specify decision options. Moreover, we can exploit them to explain our own recommendation about making decision options clear before evaluation is undertaken.

For example, Lewy identifies three decisions that might stem from an evaluation done by an evaluation unit. The first is the selection of program components: what should be taught, what materials might be included in the teaching, and so on. He is careful to point out that the program developer is ultimately the one who must choose from among several alternatives examined in evaluation. The second decision is modifying a program. He suggests that it "may turn out that some element such as exercises, illustrations, or explanations contains certain flaws". It is up to the evaluator to call attention to these. The third decision option has to do with qualifying the use of the program. Here he stresses the "optimal" conditions under which the program might work or the minimal conditions for usage. This includes, for example, whether the program will work with little or no training of teachers rather than with a good deal of training, whether equipment, space, and the like are available, and so on.

This point is important for U.S. evaluators. The Holtzman Report pointed out the inability or unwillingness of various audiences to specify decision options before an evaluation is actually undertaken. Yet here Lewy implies that the process is almost a matter of course for the Israelis. More impor-

tant, Lewy recognizes some decisions that are independent of budget and therefore less controversial.

Davis is careful to explain decisions that could emerge on the basis of his special standardized evaluation plan and to urge that these decision options be recognized before evaluation is undertaken. His options depend on the results of the evaluation. For instance, the first possible result is that the standardized evaluation yields an estimate of no effect or of negligible effects for the optimal version of the program. The implications for decisions are that one can question the "efficacy of the program in any form" and may recommend either dropping it or revising it. The second possible result is that the standardized evaluation results in a large program effect and all subsequent field evaluations show negligible program results. It is at this point that the program director, according to Davis, has to think about alternative ways of improving the field version of the program. The third case is that evaluation of the optimal version yields a large program effect and "one of the generalized evaluations results in a moderate effect". Here Davis suggests that the difference is a matter of consideration and cost benefit analysis. The final outcome of interest is the standardized and generalized evaluation result in approximately the same moderate or large program effects. It is at this point that one goes to the next step and tries to understand how the cost of programs can be reduced.

Enumerating Uses of Information. Tamir's strategy of briefly identifying the specific findings of evaluation and their eventual uses is similar in spirit, although not in detail, to the way the U.S. Department of Education Office of Evaluation's annual report is composed. The uses he identifies are interesting.

For example, he cites a 1974 paper showing low student interest in the study of botany. The subsequent action was integrating the study of topics common to both botany and zoology. He cites a finding that students had trouble applying statistics in biology with the consequent action of preparing curriculum materials on the use of statistics in biology. Another finding was that students in agriculture schools achieved poorly on this version of the material, probably because it had no applications to agriculture. Action was taken to prepare special modules to "suit the needs and interest of agricultural students". Similarly, it was found that culturally deprived students whose families come from developing countries

achieved poorly and the consequent action was preparing a Hebrew version of the material of the test designed for the original course and designing supplementary modules as well as in-service teacher training.

Such items are very persuasive. More importantly, a listing of the sort provided by Tamir might serve as a good model of the sort of information which should be recognized in the university training of evaluators. It is the kind of information that can be provided to program development staff to guide them in making decisions about how to "reason from the data" and to senior executives and perhaps legislative staff to show how evaluation results have been used. And the model is likely to be useful at the local level to illustrate use at least and to foster invention of other approaches at best.

Definition of Use. None of the seminar papers put much stress on defining the use of evaluation results. But several recognize difficulties engendered by ambiguity in the word and a variety of definitions are implicit. Blass, for instance, recognizes, as the Holtzman Report does, that arguments about utilization may be gratuitous, citing the "usual almost ritual litany about the underutilization of research results". The implicit definitions range from examination of a report, judging from the Kugelmass paper: "The very process of bringing... senior decision makers (through the Office of Chief Scientist) into continued examination of the research... may be the most important product of the process of evaluations, and not necessarily specific decisions". The Holtzman Report recognizes the same kind of use in defining evaluation and exploits it in enumerating references made to evaluations in Hearings routinely issued by Congressional Committees charged with authorizing funds for education programs.

Tamir's listing is much more specific and implies a different category of uses, also identified in the Holtzman Report: specific decisions. The category is important but was rather difficult to examine. To be sure, some evaluations, such as the National Institute of Education's Compensatory Education Study are remarkable in that evidence on use is available from the public record, such as Hearings, and can be corroborated through interviews with legislative staffers and bureaucrats. Moreover, it was relatively easy to trace ties between items in the Study and subsequent changes in federal regulation and administrative practice. Others are not as easy. Overreporting of use is likely, for instance, if one talks only to the producers of a study. Underreporting is chronic partly because of memory lapse, the absence of written accessible records on use, and the time it may

take for a report to filter through several layers of bureaucracy. The centralization of Tamir's project may make tracking utilization easier than in a decentralized enterprise.

Factors Influencing Utilization. For Blass, the most important factors influencing use are the political situation, the organizational structure of the society, personal attributes of the evaluator or of the decision maker, especially of the latter, the state of the art in the scientific discipline, and the character of the issues.

The Holtzman Report did not frame the influences on evaluation this way. But the conceptualization seems natural for Israeli operations and for some U.S. activities. My understanding is that evaluation is more centralized in Israel, if the organization of the ICC, the High School Biology Project, and the Office of the Chief Scientist are any index. This stands in contrast to many U.S. efforts in that evaluation responsibilities are dispersed across levels of government - local, state, and federal - and agencies within levels of government. Judging from the Kugelmass paper, the Office of the Chief Scientist can relate directly to the political concerns of the Ministry of Education. In the U.S., the several layers of bureaucracy between the Secretary of Education and the Evaluation unit within the Department, for good or ill, probably affect the political attentiveness of the latter and the receptivity of the former.

Centralization does appear at times in special studies undertaken in the U.S. For example, the NIE Compensatory Education Study was created by law to examine, among other things, how federal funded programs for poor primary school children fared. It was centralized in that a team approved by Congressional staff was created to be answerable to the Congress alone. It was deliberately sensitive to political issues and accommodated them through continuous negotiation between the evaluation groups and Congressional interest groups. The personal attributes of the group leader, Paul Hill, including his prior experience as a bureaucrat and Congressional staff member, seem to have been very influential in producing useful results.

There are, of course, ways to characterize factors that influence the use of evaluation other than the one that Blass proposed. In some cases, they may be more productive. Consider, for example, that each of the seminar papers stresses the richness and the diversity of the process by which evaluation results are used or not used. They are important in this respect but, for the moment, let me propose a model which may help to understand better the order that underlies the diversity.

The following questions imply distinct events that determine whether evaluation results are used:

- . Does the prospective user know about the evaluation results?
- . If results are known, are they understood?
- . If understood, are they believed?
- . If believed, does the user have the ability and willingness to use them?

When all are answered in the affirmative, then results are likely to be used. But the first negative will reduce any possibility of use, especially if the potential user is the same at each step.

This way of describing the utilization process is useful by itself. For example, it suggests that simple probabilistic models may be helpful in understanding why some research on utilization is misleading, and how one might enhance utilization. The simplest such model posits that each event is independent and a probability is attached to each question's resulting in a Yes. If the probability for each is only 1/2, say, then the overall probability of a "use" occurring is $(1/2)^4 = 1/16$. If, as I suspect, the odds are lower on each Yes, say 1/4, then the probability of a user's knowing about results is 1/4, the probability of the user's understanding them is 1/4, and so on; the overall odds against results being used are 255 to 1. Not very promising.

Other models though are more realistic. If, for example, we suppose that the evaluation process is centralized so that the prospective user is also the producer of the information, then the probability of being willing to use the information is conditional on the prior events and the likelihood of use is closer to 100%. Indeed, the conditional model is a numerical representation of what happens when a liaison person is used in ICC evaluations in Israel and when a brokerage system is used to translate findings into usable results as Cooley (1980) did in the Pittsburgh school district.

The model does suggest that we collect data at each stage, to properly estimate odds on ultimate use of evaluation results. But little work appears to have been done on this problem in the U.S. Most research deals with the question: What does the process of use look like in particular case studies? It seems to me that the case studies are important to inform such models, as well as being important as qualitative descriptors of the process. But they require large samples of events across sites or of multiple events within site to obtain decent estimates of parameters in the model.

Notes

- ¹ The research on which this paper is based was supported by the Division of Evaluation of the U.S. Department of Education and the National Institute of Education. The discussion of Israeli work here is based on material presented at the Israel-US seminars on evaluation in June 1980. The material on U.S. policy adopts heavily from Boruch and Cordray (1980) and other documents cited in the text.
- ² See Verma's (1977) discussion of research on medicine in medieval India and Rabinovitch's (1973) fine description of rabbinic thought on evidence during the ninth through twelfth centuries. The Syrian experiment is described by American researcher F.S. Chapin (1947); the English and Scottish tests are described in Cochran (1976) among others.
- ³ Two other efforts, independent of these, are worth examining because their conclusions differ at times from ours. The first, undertaken by Rand Corporation staff, appears in a monograph by Pincus (Ed.). The second, prepared by members of Stanford's Consortium on Evaluation Research, is presented in a volume by Cronbach and others (1980).

REFERENCES

- AIRASIAN, P.W., KELLAGHAN, T., and MADDAUS, G. The Consequences of Introducing Educational Testing: A Societal Experiment. Boston, Mass.: Boston College, Center for Field Research and School Services, October, 1978.
- BEN-SHAKHAR, G. and NEVO, B. Some Thoughts Regarding the Possibility of Evaluating a National Matriculation Examination System. Hebrew University, Jerusalem, March, 1981.
- BLASS, N. Educational Research and Educational Policy-Making: Reflections on the State of the Art Based on Empirical Evidence in the Israeli Ministry of Education. Office of the Chief Scientist, Ministry of Education and Culture, Jerusalem, March 1981.
- BLASS, N. Research Utilization and Scientific Knowledge Impact on Public Policy: Some Thoughts and Guidelines for Proposed Research. Office of the Chief Scientist, University of Education and Culture, Jerusalem, 1980.
- BORUCH, R.F. and CORDRAY, D.S. (Eds.) An Appraisal of Educational Program Evaluations: Federal, State, and Local Agencies. Report submitted to the U.S. Department of Education. Evanston, Illinois: Psychology Department, Northwestern University, 1980. (To be published by Cambridge University Press).
- BORUCH, R.F. and WORTMAN, P.M. Implications of Educational Evaluation for Evaluation Policy. In: D. Berliner (Ed.) *Review of Research in Education*, 1979, 7, 309-363.
- BORUCH, R.F., WORTMAN, P.M., and CORDRAY, D.S. (Eds.) *Reanalysing Program Evaluations*. San Francisco: Jossey-Bass, 1981.
- CHAPIN, F.S. *Experimental Designs in Sociological Research*. New York: Harper, 1947.
- COCHRAN, W.B. Early Developments in Comparative experimentation. In D.B. Owen (Ed.) *On the History of Probability and Statistics*. Base: Marcel Dekker, 1976, 1-27.
- COOLEY, W.W. Improving the Use of Evaluation Results. Presented at the Annual Meeting of the Evaluation Research Society, Arlington, Virginia, November 21, 1980.

- CRONBACH, L.J. and Associates. *Toward Reform of Program Evaluation*. San Francisco: Jossey-Bass, 1980.
- DAVIS, D. Standardized Evaluation of Education Programs. Hebrew University, Jerusalem, June 1980.
- HILL, P. Evaluating Education Programs for Federal Policy Makers: Lessons from the NIE Compensatory Education Study. In J. Pincus (Ed.) *Educational Evaluation in the Public Policy Setting*. Santa Monica: Rand Corporation, May 1980. (R-2502-RC).
- KENNEDY, M. School District Use of Evaluation Research. Presented at the Annual Meeting of the Evaluation Research Society, Arlington, Virginia, November 21, 1980.
- KUGELMASS, S. Considerations Toward a Policy of Evaluation Research: The Case of the Chief Scientist at the Israeli Ministry of Education. Office of the Chief Scientist, Ministry of Education and Culture, Jerusalem, 1980.
- LEWY, A. Continuous Evaluation Services Attached to Large Scale Action Projects: The Case of the Evaluation Unit at the Israel Curriculum Center. Israel Curriculum Center (ICC), Ministry of Education and Culture, Jerusalem, 1980.
- LEWY, A. Standards for Research Data. Israel Curriculum Center, Ministry of Education and Culture, Jerusalem, 1980.
- MAGIDSON, J. Toward a Causal Model for Adjusting Preexisting Differences in the Nonequivalent Control Group Situation. *Evaluation Quarterly*, 1977, 1, 399-420.
- RABINOVITCH, N.L. *Probability and Statistical Inference in Ancient and Medieval Jewish Literature*. Toronto: University of Toronto Press, 1973.
- RAIZEN, S. and ROSSI, P. (Eds.) *Program Evaluation in Education: When? How? To what ends?* Washington, D.C.: National Academy of Sciences, 1981.
- RIECKEN, H.W., BORUCH, R.F., CAMPBELL, D.T., CAPLAN, N., GLENNAN, T.K., PRATT, J., REES, A., and WILLIAMS, W. *Social Experimentation*. New York: Academic, 1974.
- ROSSI, P., BERK, R.A., and LENIHAN, K.J. *Money, Work, and Crime*. New York: Academic, 1980.
- STALFORD, C.B. NIE-Funded Studies of Local Evaluation. Presented at the Annual Meeting of the Evaluation Research Society, Arlington, Virginia Nov. 21, 1980.
- TAMIR, P. Continuous Evaluation as an Integral Part of Curriculum Development: The Case of the Evaluation of the High School Biology Project at the Israel Science Teaching Center. School of Education and Israel Science Teaching Center, Hebrew University, Jerusalem, May 1980.
- VERMA, R.L. Graeco-Arabic Medicine in Medieval India - Its Hakims and Hospitals. *Hamdard Medicus*, 1977, 20, 26-40.
- WHOLEY, J. Evaluability Assessment. In R. Rutman (Ed.) *Evaluation Research Methods: A Basic Guide*. Beverly Hills, CA.: Sage, 1977, pp. 13-37.
- ZA, I. Latent Variables in Causal Models. School of Education, Tel Aviv University, Tel Aviv, June 1980.

CONSIDERATIONS TOWARD A POLICY OF EVALUATION RESEARCH:

The Case of the Chief Scientist at the Israeli Ministry of Education

Sol Kugelmass

The Office of the Chief Scientist is a relatively new unit in a number of Ministries in Israel. The role conceptions of the office vary from ministry to ministry, but all include centralized involvement in research support and research operation. In the Ministry of Education there have been growing requests for research directed toward evaluation and accountability. This may be seen as part of an increased desire for a more scientifically based policy approach to be provided by the social sciences.

Some educational evaluation research in Israel is carried out in universities and other institutions which are outside the realm of the Chief Scientist. All of the evaluation research connected with the Ministry of Education, however, is either directly supported by or at least reviewed and influenced by the Chief Scientist. Consideration of the experience in this area during the few years of the existence of the office may be of some interest and may possibly suggest guidelines for a policy toward evaluating research within a centrally organized system of education operating with quite limited resources.

There is a large and continually growing body of literature covering many aspects of evaluation in education as well as in related social activities. While much of this literature is directed toward specific techniques or to overall contextual problems, relatively little attention has been paid to the formulation by a governing unit of a policy toward evaluation research as such. One of the probable reasons for this is that there are likely to be great differences from situation to situation. Given such situational specificity in addition to the problematic aspects of educational research in itself (see Cohen and Weiss, 1977), it is understandable why one might be sceptical of the likelihood of developing a useful overall policy. Recent experience in our office as well as informal accounts received from other

government agencies suggest that it would nevertheless be worthwhile to attempt such an exercise.

One might begin by considering the role of the Office of the Chief Scientist and its place in the Ministry. The Chief Scientist (hereafter CS) is a senior academic whose training is in one of the behavioral sciences closely related to education. His very small staff includes two other full-time social science professionals and a secretary. Since there is almost no Ministerial in-house research staff available, this compact unit operates by drawing on outside academics and professionals and by consulting with Ministry officials. The office is expected to select, monitor, and evaluate all research operations connected with the Ministry (at present forty-four projects are being monitored). In addition, the Chief Scientist is expected to serve as advisor to the heads of the Ministry and to participate in important relevant committees.

Research proposals are selected after evaluation by a research committee which has formal procedures and which utilizes relevant ministerial officials and outside scientific experts as referees. This research committee is chaired by the CS and includes, at the present time, the head of the planning department of the Ministry and four outside university social science experts. Whenever a research project is approved for support, efforts are made to suggest an advisory committee to accompany the research operation. The advisory committee usually includes a well known academic and a ministerial official who is directly involved with the project. It should be stressed that proposals placed before the research committee have first been carefully screened and to some degree reshaped by the office staff. Any evaluation policy that might be formulated will have been largely derived from experience at the decision-making level of the research committee. In turn, it will be of use in guiding the deliberations of this group in the future. It will also be of use to the CS in his future interaction with the heads of the Ministry and with the scientific community.

One of the basic desiderata for an overall policy of evaluation would, of course, be an overall plan of operation for the Ministry of Education, including clear priorities and long-range planning and commitment. This should be one of the major benefits of a centrally organized system of education. Examination of the past suggests that such an overall plan is not forthcoming. This may be due to the fact that it would be more appropriate to describe the Ministry as a centralistic rather than as a monolithic educational organizational system. Thus, even aside from shifts due

to political changes, the non-monolithic character of the system results in differences in priorities within the central units, as well as ambiguity in the formulation and understanding of goals. On the whole, almost all of the members of the Ministry leadership expect the CS to give evaluation research the highest possible priority. It is also clear that different policy-makers want these evaluations in many different areas of operation¹. While other factors would certainly be important, it seems desirable to press for some systematic overall statement of priorities by the Ministry. This would not of course replace the need for specific input from those responsible at the programmatic level. Partly on the basis of the deliberations mentioned in this paper, it was decided to make an attempt to request such a priority list from the Ministry. A formal request was recently made to appoint a committee composed of members of the Pedagogic Secretariat, the Ministerial administration and the Office of the Chief Scientist, to recommend an overall list of research and evaluation priorities to help guide the Research Committee in its selection.

A review of the recent past indicates that there have been evaluation activities that did not receive the formal attention of the Pedagogic Secretariat. Some of these activities were intimately tied with the specific projects of a particular unit on a short-term basis. The case of the Evaluation Unit at the Curriculum Center is discussed in a separate paper by Prof. Aryeh Levy. The current policy, determined by present conditions, is that short-term formative evaluation of a continuous nature need not be reviewed by the Research Committee of the CS. On the other hand a more detailed policy will have to be formulated in regard to the long-range summative evaluation of such programs. On several occasions the office of the CS was placed in the unenviable position of having to assume responsibility of a project which was already underway. A description of some of the complications of evaluation in such a case has been made by Blass of this office, in a separate paper in this volume. The general policy of the Ministry to the present has been to limit attempts to develop intramural research. Among the important factors leading to this policy were budgetary and administrative considerations, but there also seems to be doubt about the general credibility of such research results.

In this connection it is important to recall the other advisory functions of the CS. He may find himself in a situation of evaluation other than that of the chairman of the research committee. In fact he may have to serve as an interested advisor requesting formal evaluation research for a unit ope-

ration or program. An interesting and important case in point has been described by Blass. This involves the very large scale Revakha (Welfare) program started about five years ago and designed to help reduce the social gap, primarily through educational means. This is without doubt the biggest special program operating within the Ministry, with its own framework and budget. While this framework includes direction and supervision by a central governing body, it has adopted as one of its basic principles that the specific implementation of the program at the local level should be tailored to the expressed requests and choices of representatives of the local community. At the present time the program involves about eighty different communities throughout Israel.

The analysis by Blass describes the situation in 1975 when the previous CS was asked to participate actively with the leadership of the Ministry in determining the guiding policy for this program. The CS did this by drawing on a small special committee of outside academics. He also made certain specific suggestions both in terms of the goals of the program and the means to reach them. As Blass reports, this led to the CS becoming involved as a policy-maker, and drew him into conflict with others at that level who challenged the role of the social sciences advisor in that capacity. Blass concludes that while the program adopted some of the suggestions of the CS related to the choice of means, there was strong reservation about his determination of basic policy.

Had he continued in an important role in basic policy determination, the CS would have faced problems as chairman of the Research Committee determining evaluation research policy for this same program. The present CS has attempted to minimize his advisory role to the program in the area of basic policy determination. He still participates as one of three scientific advisors to a special unit created to be of assistance to the directors of the program, but in contrast to the earlier role, he now advises the director of the program itself rather than the Director General. Discussions of and arrangements for evaluation of this program have thus shifted into a situation where the CS is seen as an in-program advisor on the one hand, and an agent for technical assistance in obtaining the necessary tools to be used in different kinds of evaluation on the other. This part of the case history suggests that the CS must be aware of certain role conflicts which will influence his ability to function as the central and main agent responsible for evaluation in the Ministry.

Other important lessons emerge from a review of this large program. As

the program moved into action, increasing pressure for a system of evaluation appeared at the level of the Ministry leadership as well as within the program itself. At a later date pressure for evaluation even emerged at an unexpected level - that of the local operation. In responding to the initial pressure - particularly to the pressure of the Ministry leadership for some form of professional evaluation, the previous CS arranged for a visiting evaluation expert to make a fast survey of one of the local projects. This particular project turned out poorly not only because the Ministerial project directors resented certain aspects of the assignment procedure, and objected to some of the resulting criticism contained in the report, but also because the report was viewed as being superficial and unworthy of a serious social-science approach.

At this point it is important to state that the CS faced a rather difficult situation, in that there were few appropriate scientific forces available for such an evaluation assignment. Empirical educational research in Israel is performed for the most part by the universities or university affiliated institutions. Only quite recently has there been some shift in emphasis toward more applied and specific task-oriented research. For a number of reasons related to the sociology of Israeli university hiring and promotion procedures, it was not easy to attract university academics to this kind of evaluation research. While there has been some change for the better in this situation, this is a fundamental problem for any evaluation policy and should be given further consideration.

The ministerial and project leadership felt that first priority should be given to some kind of survey of the operations being carried out in the field. It was also hoped that the universities would be able to supply appropriate scientific leadership in carrying out more intensive programs of evaluation for the local programs in their own particular vicinities. Another approach was to evaluate specific programs that were supported in a number of the local communities. After much effort, it became clear to the CS that the limited university interest could be more easily attracted to the second and third kinds of evaluation projects. The CS thus thought it appropriate to press the top-priority survey upon a non-university research institute which had loose ties with the Ministry of Education. In spite of major efforts to coordinate this operation with the program directors and help from a special advisory committee, this project had to be stopped. This was a very difficult decision to be made by the new CS with the help of the research committee. It was the first, and until the present, the only case of an invited research

project being terminated by the Ministry.

Both of these unfortunate attempts to respond to the pressures for evaluation indicate that another basic element in an overall policy toward evaluation must be to avoid such hasty responses in the future. It is wiser to avoid evaluation if the proper conditions cannot be met, than to pay the heavy price of inadequate or poor research that results when the conditions do not permit proper evaluation. In order to minimize the pressures, it would seem necessary to have the CS invest considerably in attempting to help the policy-makers become more aware of the necessary and realistic limitations of his efforts. This presents a major dilemma, in that the CS must find a way to consolidate the growing interest in evaluation with a scientifically assisted policy approach that will be realistic and without oversell, yet without "turning off" this interest by failing to produce all the products desired by the educational leadership².

A review of the past three years indicates certain progress, but no agreed upon fundamental method for research into the Revakha program. Research projects for the evaluation of several of the specific programs were developed and passed through the channels of the Research Committee. One of these has already been completed, and another is close to its final stage. There was less success in operations at the local level, but even here there were some very interesting results reported from Kiryat Shmoneh by Schlesinger. Nevo's paper is an analysis of one of these cases. Without doubt the biggest problem is how to evaluate the program as a whole.

Following the failures mentioned above in obtaining outside evaluation assistance, it was decided to establish a small unit attached to the central administration of the program. This small unit is attempting to gather and organize basic information on activities at the local level. It is also moving to some extent into formative evaluation with feedback to the local community operation. It is hoped that the basic descriptive data will be of use to any future researcher attempting to do more systematic, objective, and summative research. Another point to be considered for the overall policy of the CS is that it may not be possible to avoid setting up the kind of specialized internal unit required for information and formative evaluation.

Consideration of future plans for summative evaluation within this large program has already suggested that such efforts will be problematic because of the lack of a proper information base and technical infrastructure.

Some perspective may be obtained from a relatively recent interim report

on the NIE Compensatory Education Study by Hill and Rothberg (1976). When this important study was initiated, "there were no current, nationally representative data on the participants, the services offered to them, or the costs associated with compensatory education projects and programs". A study which reviewed state reports (Gamel, 1975) was critical of the quality of information they provided, and it was thus concluded that these state reports could not be used as the basis for NIE's research on compensatory services delivered by school districts. The authors also describe the limitations of assessments by non-governmental independent groups, as well as the early nation-wide and later more focused evaluations. One of the first projects commissioned, therefore, was a national survey to collect information on Title I and state compensatory education services delivered by school districts. Another frank and useful footnote in this report seems very relevant:

NIE's approach owes much to earlier evaluators and builds on their experience. The history of Title I evaluation has been a slow process of learning how to live with the realities of the program. Evaluators have learned through experience that the diversity of program operations in different school districts rules out "black box" methods of research, with focus on inputs and outputs without consideration of what the program is in practice. They also have learned that the data routinely supplied by LEAs (local education authorities) are not of uniformly high quality and cannot support the kind of rigorous analysis required in national evaluations.

It should be pointed out that the kinds of problems mentioned by Hill and Rothberg are probably even greater in the Israeli program which is much more complex, heterogeneous and less developed in administrative practice.

Pressure from local units to carry out evaluation studies of their own oriented toward school achievement uncovered another limitation. Perhaps even more than in other research oriented countries, there is a lack of standardized achievement tests available in Israel. A report by R. Rogel as well as informal examinations of the office of the CS suggest that many problems exist in the system of examination advisors used within the Ministry. This suggests that a national effort to develop and utilize appropriate tests in the school system must be considered seriously. At the moment a feasibility study of such an undertaking is under way, and the report should be available in a few months. A recent report by Hill and Rothberg (1977) describes efforts to develop a national testing system in the U.S. which

might use achievement test scores to allocate Title I funds. This suggests that an overall evaluation policy of the CS must consider other related research and development plans.

Perhaps it may be useful to distinguish between the specifics of organization and structure, and the fundamental issue of the feasibility of a good summative evaluation of a large-scale heterogeneous educational program. Some specific suggestions for better utilization of the scientific manpower at the universities in educational evaluation have already been raised in this seminar. They include better communication between the office of the CS and certain departments, approaching research groups rather than individuals, and even long-range support for centers of educational evaluation. Other suggestions for the development of the necessary infrastructure have been mentioned above. It is important that this be presented as part of an evaluation policy which is assimilated by the leadership of the Ministry. On the other hand it may be that the CS will have to take on a more active role in considering the limitations of some kinds of evaluation. Naturally he should work together with the relevant experts, but he may have to take on a central role as their spokesman if consensus is reached. Again, some preliminary discussion in this seminar has been given to possible approaches to the large-scale summative research, but there does not yet appear to be an agreed-upon approach which seems feasible right now. There seems to be debate about the desirability of even partial evaluation approaches, and these might not satisfy the demands expected in the future on a higher political level.

The gap between the expectancies of the decision makers and the evaluation program of operation may be noted in detail in another large scale evaluation of the Reform - the junior high school change described by Blass. In 1968, the Israeli Knesset legislated a major structural change which instituted a new *stix*, three, three year sequence in government schools, designed to replace the existing eight year primary school and four year high school program. The new three year intermediate section would serve as a larger physical unit, in which it would be possible to carry out a wider program of integration of the different social groups of Israeli Jewish children. One of the recommendations of the Knesset committee was for research to accompany this new program in order to evaluate and improve it. Three years after the program started, a team of well-known educational researchers was approached, and the research initiated.

It is of interest that the researchers were given a rather free hand in the design of the evaluation research. Only about half a year after the

research operation had started, was there a formal meeting with the administration of the Ministry of Education. During the several meetings that took place during the next two years it became clear that there was some disappointment within the Ministry. Part of this may have been due to the fact that it would be nearly impossible to satisfy all of the different views expressed by the different members of the administration. It also seems that most of them were interested in what may be termed formative or process evaluation that could be used to actively influence the educational program as feedback. There even seems to have been some disappointment among Ministry officials about their role as an advisory committee to the research team.

During further discussion within the Ministry, following the presentation of the preliminary report there were other expressions of disappointment about aspects of the program not covered by the research. Much of this would seem to be unfair, and the question of interest to us would be to what extent this disappointment and misunderstanding could have been avoided. In this case, it seems that more extended and detailed negotiations between the evaluating researchers and the decision makers might have been useful. At least some of the disappointment could have been minimized if the distinction between formative and summative evaluation had been considered. The particular longitudinal design decided upon by the scientists had certain advantages, but there seem to have been disadvantages that were not understood. Another point that is of interest is that the advisory committee that met with the research team consisted solely of members of the Ministry. It might have been of some use if other non-involved professionals, who might have helped close the gap in perception between the researchers and members of the Ministry, were added. The most important conclusion for our evaluation policy is that attempts to clarify the goals and procedures of any evaluation research should be made as much as possible beforehand. The preparatory stage should also include the specification of the particular forms of communication and negotiation between the evaluators and the decision-makers during the research operation. Even if strong differences of opinion and expectancy turn up during this preliminary period, it will be possible to suggest a number of different approaches, including the substitution of different teams. Given the importance of this major structural change in the Israeli educational system, it is quite striking that little or no systematic formative research apparatus was established. Another look at the report on Title I evaluation illustrates this point very clearly in terms of goal specification. "NIE's strategy for assessing compensatory education programs begins with the recognition that the program has several purposes. It differs from the early

national evaluations of Title I which considered only the third fundamental purpose...". It was thus years later that a more serious consideration of goal specificity was undertaken by Congress and the evaluators.

A more dramatic, if small-scale case in point of the necessity for prior agreement between the evaluating researchers and the decision makers may be noted in the recently completed research on mother-child game clubs. This program had been adopted by a number of local communities in the previously mentioned large-scale Revakha program with a great deal of enthusiasm. Only after the essentially negative results were reported were there strong complaints that the goals of the research had not been sufficiently considered, and that the expectancies had been unrealistic. It then turned out that one of the departments in the Ministry dealing specifically with this topic had not been consulted. In this particular case a great deal of effort was required on the part of the office of the CS to prevent the negative findings from being ignored totally.

Based on this and other cases attempts to maximize pre-research negotiation and specification have increased. These attempts have turned up a considerable number of problems, and have produced quite a bit of frustration. There is no doubt that this procedure often leads to delay. In cases of great urgency, such as the felt need for information about drugs or violence, this is particularly problematic. On the other hand, such negotiation may lead to the conclusion, as in the present case of research on violence, that the decision-maker really could not have made much use of the originally designed evaluation research which had almost been agreed upon. It appears that such negotiations are a difficult thing for decision makers, and often place them in a strong anxiety-producing situation. It is also quite frustrating for some of the researchers who have their own schedules and balances of resources to consider. Too much pressure in this direction can be counter-productive, but at this stage of our experience it appears that insufficient attention has been given to the benefits of pre-research specification involving both the researchers and the clients.

The discussion of pre-research negotiation and specification has been primarily on the general goal level. Actually there is reason to believe that there may be many advantages in moving to the level of methodology and data analysis as well. Continued negotiation and consideration between researchers and the Research Committee of their ability to draw conclusions led to renewed negotiations with the relevant Ministry officials in a very large scale attempt to evaluate the role of non-formal education. Because

of methodological constraints found to be inherent in the original design, it was decided to change the research program. It was deemed more useful to undertake a broader approach which would include many more factors, even if this approach could not provide the original clear cause and effect analysis originally entertained. This required further negotiation with decision-makers and led to a consortium of interested officials from three different Ministries. The point to be stressed is that by making the methodological implications clear to both researcher and client, the Research Committee was able to change the form of the research so as to make it of interest to other clients. Of course a heavy investment of energy and a large amount of frustration was involved in all of these negotiations which almost blew up several times. While there is no objective basis for concluding that this research will be a success, it seems fair to say that there was general consensus that the research product will be technically more sound, and will probably mean increased involvement by the clients.

A certain amount of negotiation will continue to take place during the research operation, and possibly afterward as well. As mentioned earlier, an advisory committee is usually appointed by the CS to accompany all large research or evaluation projects approved for funding. Depending on the particular situation and the problems that may arise, the committee may be quite active in helping to decide about methodological choices, data interpretation, or even requests for more time or funding. Since this committee includes both researchers and at least one involved decision-maker, it provides for continued discussion and understanding about the research, throughout the process. The chairman of the advisory committee also maintains contact with the Office of the CS in order to keep it informed of progress as well as problems.

One of the most important functions of the Office of the CS is to review and evaluate all research undertaken by the Ministry after it has been completed. Experience has shown that by having active and communicating advisory committees, such reviews will be less likely to be critical or controversial. An interesting case in point would be to contrast the situation in two very large and important projects. One of these, the Van Leer study of the Israeli primary school, was started in 1971 and funded by the Van Leer Foundation. Only at a later date were supplementary funds requested from the Ministry of Education. This study did not, therefore, go through the usual procedures. In 1977, however, it was evaluated by the Chief Scientist, who reported his conclusions to the Ministry and to

the scientific community at large. The results of this study received a great deal of attention in the press and radio. It was also discussed in the education committee of the Knesset. During these discussions, the critical review by the Chief Scientist, which was relatively mild, received quite a bit of attention. Ongoing discussions in an advisory committee would probably have reduced the necessity for a critical evaluation, and would, in any case, have put the criticisms into a better perspective.

This did, indeed, take place in the case of the evaluation research dealing with the Reform (junior high school change) previously mentioned. Some of the problems concerning the use of an all-ministerial committee have already been mentioned. Three years after the research had begun a different advisory committee, including outside academics, was appointed by the CS. A couple of years later, following various delays in the appearance of the final report, this committee was strengthened. This change was deemed necessary following reports in the newspapers that suggested that the Ministry was delaying the appearance of the report because of its findings. The advisory committee worked intensively with the research team and, while the researchers did not accept all of the specific suggestions, they were able to present a better conceived and balanced report. In this case, the appearance of the researchers, the advisory committee and the CS before the Knesset committee of education was more coherently tied to the issues of the report. While there were clearly some differences in position among the scientists, these differences were presented to the Knesset committee members as legitimate professional views.

Up to now our analysis has focused on the need for pre-, ongoing, and post-research interaction between the researchers and the decision-makers, with the CS as an active intermediary at one stage and an evaluator of the research at another. It was suggested that judicious handling of this interaction may lead to better and more relevant research with fewer conflicts in the research presentation. It would appear that such treatment would also maximize the possibility of research and evaluation utilization.

In many ways, the problem of research presentation and utilization may be regarded as the most serious problem of evaluation. Many authors have stressed different aspects of this problem, and as Alkin and Doillak (1979) have recently noted, "the accounts (are) so gloomy that one sometimes wonders whether effective evaluation is ever possible". These authors did not, however, despair and have made an attempt to examine the evaluation

utilization process systematically through empirical study. Our own experience in Israel would, on the whole, tend to support the pessimistic view of evaluation utilization. On the other hand there have been indications that it may be possible to facilitate utilization. One Israeli experience may be of interest, in that a serious attempt was made to do so.

The Van Leer study of primary schools is probably the largest serious empirical research study in education undertaken in Israel. In many ways it may be regarded as similar to the well known Coleman report of the U.S., both of which may be regarded as primarily descriptive rather than as formal evaluation efforts. The results of the Van Leer study were used by different people to draw conclusions about a number of issues, including such controversial ones as integration, the religious vs. the non-religious school systems, etc. These issues were discussed in the press, on the radio and in the Knesset as well as in academic and professional circles. The CS and the Ministry leadership felt that steps should be taken to utilize the research in a more systematic and long-range operational way. It was therefore decided to appoint a committee consisting of the CS, the head of the government religious school system, the head of the Ministry curriculum center and the head of the secular primary school system, the last to act as chairman. This committee was given a special budget for a full-time psychologist coordinator and was asked to attempt to utilize as fully as possible findings of the study for the Ministry. It was decided to appoint six sub-committees of appropriate educational professionals drawn from different supervisory and field levels. Five of the sub-committees were to examine the results in terms of specific subject-matter and the related curricula, and the sixth was to relate to the flow of resources to the different schools in the system. The sub-committees were given orientation to their task by the coordinator, the CS and the researchers of the Van Leer Study. All of the sub-committees were requested to make detailed evaluations of their areas in the research and to make specific recommendations to the main committee. In the case of the sixth committee which dealt with the distribution of resources, it was felt necessary to carry out a follow-up research study so as to make the data more relevant.

All of this material was transmitted to the main committee, which then submitted a detailed set of recommendations to the Pedagogical Secretariat for discussion and implementation. While it is difficult to report on specific operational changes in the field, informal information suggests that there has been some impact. In a recent discussion with the new head

of the Pedagogical Secretariat the writer was told of its impact on curricula construction.

It should be added that from personal observation the writer has noted that the formal results expressed in the reports and recommendations may not be the only or even the most important indices of utilization. The very process of having these senior decision makers examine the research in a detailed way may be the most important product. Thus the updated research report of the sixth committee written by O. Razieli has been mentioned as an important datum in a number of discussions of ministerial policy, including the most recent budgetary allocations.

Alkin and Daillak's research data suggested an analytic framework for considering the most important facets of evaluation situations in terms of utilization. Some aspects of the Van Leer research utilization may be viewed within this framework, e.g. those categories dealing with "orientation of the users" or "organizational factors". Alkin and Daillak's analysis, however, did not include the perspective for an agent such as the Office of the CS which would play an intermediary role between the evaluator (researcher) and the project administrator.

From the point of view of the Office of the CS this suggests that it is possible to maximize research and evaluation under certain conditions. It suggests that above and beyond taking an active role in the research related negotiations at all stages, the CS may have to help create particular organizational mechanisms to help maximize the research utilization. In terms of the resources available to the CS, this may suggest that even more emphasis should be given to the problem of research utilization.

Another large-scale summative evaluation study was carried out within the framework of the International Association of Educational Achievements (IEA) (Lewy et al, 1978).

Data on educational achievements were collected in 21 different educational systems. The results of this study made it possible for the first time to compare the achievement level of Israeli learners with that of learners in other countries. A cross-national comparison of results revealed that in the fields of English as a Foreign Language and Civics Israel ranked higher in comparison with other participating nations than it did in Reading Comprehension. The average achievement level of the disadvantaged children in Israel surpassed the national average of developing countries. It corresponded to the lower level of the national averages in developed countries.

The findings of this study have been used particularly by teams of the Curriculum Center of the Ministry of Education and Culture in their preparation of instructional materials for various types of schools.

Up to now we have considered some broad but very basic issues facing the office of the CS in connection with research and evaluation. It may be worthwhile to add some thought about more specific and even technical questions that have arisen in our experience.

This level of policy is closer to the actual decision-making operations of the research committee. Let us examine a few examples which seem salient. It may be that a more systematic review would turn up others. Some of these have appeared in the literature, but seem worth mentioning as they appear as central problems in our deliberations. There does not, for example, appear to be sufficient awareness of the distinction between formative and summative evaluations. This is not only true of the clients but of the researchers themselves. One specific issue in this much-discussed distinction is at what point in the development of a program one should consider summative evaluation. Our own experience has been that for various reasons it is often undertaken too early, in the sense that the program may not be sufficiently crystallized, and yet with insufficient development at that time of the appropriate tools and information base. Many of our recent efforts have been to attempt to convince both clients and researchers to focus on formative research, even if only informally, until the proper conditions permit a meaningful summative research. This policy decision is particularly important during a situation of limited budgetary resources.

Given resource (financial, manpower, etc.) limitations, it also seems important to consider the possible contribution of the particular evaluation to other programs. Unless there are other good reasons to do so, we feel that the limited resources available should not go to the evaluation of many specific programs. Instead, efforts should be directed toward better exploitation of some of the larger evaluation studies completed. During the past few years, for example, priority was given to low-cost (Ph.D.) studies relating to or actual utilizing the data from the large study on the Reform (junior high school).

On a number of occasions during the discussions of the research committee consideration was given to the most appropriate form of a control group for comparison with the treatment group to be evaluated. The writer's impression is that, given the possibility (which is unlikely at present) of sufficient

funds being available, the consensus is for attempting to compare the specific treatment with any other reasonably similar treatments rather than to only a non-treated control group. It would, on the other hand, be advisable not to use too many group variations if such would require the number of subjects in each group to be significantly reduced.

During the past year there has been a noticeable trend toward a greater focus on the independent variables in research evaluation. Previously, most questions and attention were directed to the dependent variables. Considerations of various evaluation projects proposed to deal with the comprehensive high school have for example, indicated how little was actually known about this type of school in actual operation. It became clear that a preliminary research specification of this educational system would be required before any comparative evaluation would be meaningful. Other researchers have started to consider finer analyses of the way teachers actually interpret and operate a given curriculum or program, and the possible changes in the program over time. Perhaps it is worthwhile to mention why this is an issue of policy. Along with the consideration of the priority of a project in terms of its value for the specific goals it sets out to meet, one must consider the contribution or potential contribution it may make to our general state of knowledge in this field. In the past such considerations often focussed on the "technology" for assessing a dependent variable such as the development of measures of creativity (Gur), or class atmosphere (Raviv).

The last topic might be the consideration of "non-research" forms of evaluation. It has been our experience that there are some situations which require rapid decisions which cannot wait for the more standard research approach. The CS may provide the decision-makers with an analysis by a group of experts which offers a different perspective which, while lacking the strengths of empirical research, has other advantages, such as a broader general knowledge base and increased objectivity. A very interesting situation worth detailed analysis was the handling of the Reform research by the committee appointed by the Minister of Education.

Notes

- ¹ See Discussion by H.E. Freeman on p. 16 of *Evaluating Social Action Projects: Principles, Methodological Aspects and Selected Examples*, Paris, UNESCO, 1980.
- ² See p. 317 of Boruch and Wortman's chapter in Berliner D. (Ed.) *Review of Research in Education*, Vol. 7, Washington, D.C.: American Educational Research Association, 1979.

REFERENCES

- ALKIN M.C., DOILLAK R. and WHITE P. *Using Evaluation: Does Evaluation Make a Difference?* Beverly Hills: Sage Publ., 1979.
- BLASS, N. The Chief Scientist: His Role and Image. Unpublished Master's Thesis, Jerusalem: Hebrew University, 1976.
- COHEN K. and WEISS J. Social Policy: Schools and Race. *The Educational Forum*, May 1977, 41.
- GAMEL, Nona N.; TALMADGE Kasten G.; WOOD, Christine T. and BINKLEY, Joanne L. *State ESEA Title I Reports: Review and Analysis of Past Reports and Development of a Model Reporting System and Format*. Bethesda, Maryland: RMC Research Corporation, 1975.
- GUR, Amos and ALEXANDER, Gad. Investigating the Efficiency of Activities for Fostering Creative Thinking among Disadvantaged Youths. Jerusalem: The Institute for Investigating Fostering Care, Hebrew University (forthcoming).
- HILL, Paul and ROTHBERG, Iris. *Evaluating Compensatory Education: An Interim Report on the NIE Compensatory study*. 1976, Washington, D.C. Department of Health, Education and Welfare.
- HILL, Paul and ROTHBERG, Iris. *Compensatory Education Services*. 1977, Washington, D.C. Department of Health, Education and Welfare.
- LEWY, Arie; RAPAPORT, Chanan and MERON, Mordechai. *Educational Achievement in Israel*. Tel-Aviv, Ramoth, 1978.
- MINKOWITZ, A.; DAVIS, D. and BASHI, Y. *Evaluation of the Educational Achievements of the Elementary School in Israel*. (The "Van Leer Study"). Jerusalem Hebrew University, School of Education, 1977 (Hebrew).
- RAVIV, Amiram. Giving Feedback to Teachers Regarding the Social Climate in their Classes and its Effect (forthcoming).
- SCHLESINGER, Haim. *The Other Side of the Coin: Additional Aspects of the High School Welfare Program in Kiryat Shmoneh*. Kiryat Shmoneh: Educational Section, The Unit for Coordinating Foster Aid and Welfare in the Ministry of Education, March 1979.
- ROGEL, Rene. *Activating Educational Evaluation within the Educational Welfare Program*. Paper presented to the Unit for Coordinating Foster Aid and Welfare in the Ministry of Education.

RESEARCH AND POLICY- MAKING IN THE ISRAEL EDUCATION SYSTEM A CASE STUDY

Nachum Blass

INTRODUCTION

This case study is one of a series which describes the "politics of research" rather than the topic of study itself, focusing on various aspects of interaction among researchers, administrators, politicians and the general public. Interaction of this type is largely responsible for determination of the utility and potential contribution of research projects.

The research method employed is essentially historical in nature and the report narrative in style. Attention is drawn to several particular features of the case at hand which, in combination with the specific problems dealt with in this study and those pertaining to the utility of research in general, convinced us that this topic would generate some interest among scholars in a variety of disciplines, for a number of reasons:

a) One of the most frequently employed criteria for determining the probability that a particular study will indeed be utilized is the degree of contact between researchers and their sponsors. The case at hand is considered large-scale (by Israeli standards) and such contact was continuous and at a rather high level. Relations among the parties were virtually exemplary: close ties among researchers, political figures and administrators were coupled with a high degree of academic freedom. It is most interesting to note the effects of such good relations upon the outcome of the study itself.

b) The research was directed by steering committees right from the outset. These teams - particularly the final one - assumed highly active roles, a most unusual phenomenon in the academic world. It appears worthwhile to examine the dynamics of these committees and their relationships with researchers and administrators alike.

c) This study reflects all the methodological, educational and political problems inherent in large-scale social project evaluation, namely ambiguity of goals evaluated, sample selection difficulties, measurement of tool weakness, duration, communicability, etc.

BACKGROUND MATERIAL FOR RESEARCH ON ISRAEL'S JUNIOR HIGH SCHOOLS

On 1 July 1968 the Reform in Israel's educational system (hereinafter: the Reform) came into effect. Thus, in a decisive majority vote, the Knesset (Parliament) put an end (at least temporarily) to one of the longest and most fundamental public debates concerning education in Israel. The Knesset decision marked a crowning achievement for the then Minister of Education, the late Zalman Aranne, who had exerted considerable efforts in instituting changes in both the organization and content of the educational services in Israel.

According to the late Minister, the objectives of the Reform were:

- a) The more efficient exploitation of the latent intellectual and individual potential of the younger generation.
- b) Speeding up the pace of narrowing the educational gaps among ethnic groups, particularly those between children of parents from Islamic (also referred to as Oriental) countries and those of other backgrounds.
- c) Raising educational standards in accordance with social and national imperatives.
- d) Adapting studies to suit the pupils' abilities and talents and encouraging development of their individual qualities.
- e) Providing authoritative guidance for pupils throughout their secondary studies¹.

These goals were to be achieved through enactment of a Compulsory Education Law for grades 9 and 10, separation of grades 7 and 8 from the elementary school and creation of a 7th-to-9th grade junior high school, raising professional standards for junior high school teachers, bringing together pupils from various socio-demographic regions in junior high schools, improving guidance services, etc.

The main thrust of the objections to the Minister's suggestions came from the ranks of the Israel Teachers' Union, headed by Dr. Shalom Levin. The principal contention was that elementary education had already proven itself

and ought not be ruined by severing grades 7 and 8 from it, and there was no proof that the age of 12 was indeed suitable for a pupil's transfer to a different educational system. If far-reaching organizational change was desired, it should have been based upon empirical research and longitudinal follow-up studies. When individualized instruction is conducted and wide gaps exist, the result is often segregation by ability level within classes. Hence such gaps should be eliminated while pupils are still young - that is, the elementary school system should be reinforced. It was doubtful whether the grouping of adolescents would aid those with learning disabilities to progress, but it was certain that social damage would result. In summary, while agreeing essentially with the basic objectives expressed by Minister Aranne, Dr. Levin declared:

Elementary school education may be improved through successful implementation of the reform proposed by the Teachers' Union, namely: acknowledgement of the fact that educational lags and gaps appear at an early age and this requires the concentration of efforts upon very young children; the institution of more effective teaching methods; the provision of good teachers with small classes; the development of a suitable curriculum for grades 7 and 8, the adoption of the positive aspects of the ungraded school currently being worked upon in the United States.²

As indicated above, the Knesset voted in favor of the Minister's proposal and Israel's first junior high schools commenced operation at the beginning of the 1968/9 school year. By 1970/1, when research on junior high schools was first commissioned, there were sixty such schools, and members of the first junior high school class had already completed their studies within the new system.

The members of the research team described the outset of their work as follows:

The Knesset Committee headed by Dr. Rimalt recommended that implementation of the Reform be investigated and evaluated, in order to derive appropriate conclusions and draw up suggestions for improvement. By the 1970/71 school year, there were more than sixty junior high schools in operation throughout Israel and educational researchers from the Universities in Jerusalem and Tel Aviv were commissioned to undertake this Reform evaluation project. Research commenced in 1971/2 and was financed throughout by the Ministry of Education.

At the initial stage the researchers arrived at two basic decisions:

- a) The research would take the form of a follow-up study; that is, a group of pupils would be investigated continuously throughout their three years of junior high school studies.
- b) The work would be divided between the Hebrew University team headed by Prof. H. Adler and the Tel Aviv University team headed by Profs. M. Chen and A. Lewy. The former group would concentrate upon adults connected with the Reform (i.e. parents and teachers), while the latter would be concerned with the pupils themselves. This division of labor would enable each research team to work independently (insofar as sampling, etc. was concerned), with data to be combined and rendered uniform prior to the final stages of research.¹

The researchers were first approached through the initiative of the then Minister of Education, the late Yigal Allon.

It is clear that the study was not commissioned primarily because of the Rimalt Committee's recommendations, but rather because those on the political level sought to moderate the dispute with opponents of the Reform, which had raged on despite all Knesset decisions, and even after implementation had commenced. The commissioning of an "objective and long-range" scientific study was thus an excellent method of postponing consideration of a number of issues, including the justification of the Knesset's decisions, the success of the Reform and continuation of the massive investments required for implementation². In addition, however, it must be said that Ministry officials truly did wish to learn more about the Reform's initial results.

When the research was first commissioned, the Reform was being implemented virtually exclusively in development towns* and in schools which were only in their initial stages of development, thus rendering the establishment of junior high schools relatively simple. In large cities, the Reform was carried out at a much slower pace, with the exception of Haifa - whose Oriental population is proportionately lower than that of other Israeli cities. "During the first three years, the possibilities of establishing standard junior high schools (with at least four classes for each grade) without resorting to extensive construction programs, were virtually exploited to the fullest."³

* The term "development town" describes a small urban settlement, generally located on the outskirts of a large metropolitan area or in a frontier region, populated primarily by immigrants, the majority of whom are of Oriental (i.e. Asian/African) origin.

As the air still reverberated with the echoes of the great dispute concerning the efficacy of the Reform, senior Ministry officials were concerned with other problems. Instead of academic and theoretical debate concerning projected advantages and disadvantages of the educational reform, the Ministry now faced the day-to-day problems of construction and development, including such questions as:

a) Where should new junior high schools be established? This question included a number of corollaries: should schools be set up in places where it was relatively simple to do so and authorities would readily consent, or in more difficult locations wherein the very principles of the Reform could take effect? How were land shortage problems to be overcome? What about questions of distance? Of what significance was the housing shortage insofar as elementary education was concerned, etc. ?

b) Where would junior high school teachers be obtained? If there was a shortage, would the pace of reform implementation be slowed down or was it preferable to be more flexible in terms of teacher training and level?

c) There was a shortage of textbooks, curricula and teaching methods suited for the junior high school's heterogeneous student body, particularly the weaker strata. How was this to be overcome?

d) How should the weaker strata (Group C* and the slow class) be dealt with?

e) What should be the content of the second half of the school day, which was supposed to be devoted to extra-curricular activities?

f) The continued rivalry between the two teachers' organizations was expressed both in the division of authority and in the interpersonal relations between the respective principals of the six-year comprehensive school and the junior high school which was part of it. What steps needed to be taken to prevent this rivalry from affecting the Reform?

g) What was to be the nature of government control and responsibility for junior high school teachers?⁶

* The students in regular schools are divided into four groups. Groups A, B and C are all normal groups, with A being the most academic. Below these are the "slow classes" for hard-core problematic students.

The researchers were given carte blanche with respect to the study's format and methods of implementation. They thus had to decide upon the directions of research: Would they address themselves toward responding to essential questions underlying the debate over the Reform - questions which had already been answered in a political manner - or consider the more serious educational and organizational problems which were encountered in its implementation? Their unambiguous decision was reached after internal deliberation and probably after consultation with Ministry of Education staff*, or, possibly because of a desire to satisfy the demands of the study's sponsors. In a document dated 31 October 1971, the first official and well-organized proposal for investigation of the Reform, the researchers declared:

The research will focus upon three main topics:

1. Evaluation of learning achievements in four subjects: Hebrew, arithmetic, English and science, comparing junior high school pupils with those in other types of schools.
2. Evaluation of the extent and intensity of ethnic-group social integration regarding both behavior (communication among groups) and attitudes (i.e. opinions and beliefs) among junior high school pupils as compared with others.
3. What are the opinions of parents and teachers regarding the methods employed in the education of their children and pupils? Comparing the views of adults whose children/pupils study at schools where the Reform is in effect with those of an appropriate control group.⁶

This same document also provides a general outline of the proposed sample, suggested ways of implementing the research and the scope and extent of financing required.

* Ministry involvement in guiding research is mentioned in a letter by Minister Allon to the editors of *Haaretz*, published 18 September 1971, which included the following statement: "...In addition, we commissioned a follow-up study to be carried out by university personnel, in order to accompany implementation of the Reform, investigate its achievements in the area of social integration, and determine its contribution toward the level of achievement of certain groups of pupils. The master plan, together with research findings, will provide us with data and evaluative methods which will enable optimal implementation of the Reform from the very beginning, considering the particular conditions of each applicable location."⁷

It should be indicated that the problem of sampling had become critical even at this early stage of research, as junior high schools established up to that time did not constitute a truly representative sample of the overall educational system in Israel. This fact created two essential difficulties for researchers:

A. Should the research sample represent the current situation in junior high schools at the time the study was undertaken or should it mirror the projected reality - i.e. the Israeli educational system as a whole?

B. How could they ensure that the control group matched the experimental group in all parameters, even though the Reform was first put into effect in areas with special demographic characteristics?

With regard to the first problem, the researchers concluded that the current, ongoing situation was to be described in their study. The second difficulty could be overcome, in their opinion, through the employment of various statistical techniques.

Another important decision was to include in the study one group of pupils which had been investigated previously by Professors Chen and Lewy⁹ as they progressed from fourth through sixth grades. This would provide the advantage of a single, longitudinal follow-up of one group from grades 4 through 9 (in fact, this follow-up was continued through grade 12). Even though this led to a situation wherein there was a maximum of longitudinal data available when the report was presented, this data was not fully utilized. Due to the time pressure, the decision to use the longitudinal study of Chen and Lewy did not allow for the possibility of administering a proper pre-test to the pupils at the end of grade 6 and the first tests were thus deferred until the end of the 7th grade.

The above decisions and others pertaining to the structure of research, the nature of questionnaires, etc., were agreed upon by the research workers and approved by the Ministry of Education, without their having been subjected to any comprehensive scientific investigation, and the research commenced during the 1971/2 school year.

The Ministry of education had apparently not undertaken sufficient advance preparation for the research commissioned, and the very small amount of time that elapsed between the Ministry's appointment of a study team and the commencement of field work at the beginning of the school year hindered the research. Prof. Lewy, then serving as Science Adviser to the Ministry, was

consulted by the researchers who were interested in having him on their team, but Ministry officials had not consulted with him at all regarding the scientific aspects of the program, investing him exclusively with administrative responsibility as regarded the contract to be drawn up. The Pedagogical Secretariat was likewise left out of the picture insofar as research planning, implementation and follow-up were concerned.

The research commenced in December 1971; by the end of the school year (June 1972) tests had been prepared and administered in English, social studies, science, and arithmetic; and personal questionnaires had been distributed among pupils. In addition, the following issues were investigated in a preliminary manner:

Did sixth-graders drop out en route to junior high schools? Were the results of integration significant? What were the attitudes of teachers and principals toward junior high schools? What type of teacher or principal (elementary or secondary) was most suitable for junior high schools? At the end of the school year (4 July 1972), the researchers first met with Ministry officials. Dr. H. Adler was first to speak at this meeting, emphasizing the significance of continuous reporting to the Ministry:

We have concluded that regular reports must be submitted, in light of the importance of this research project to the Ministry. We are investigating the junior high school system. When we were asked to conduct this follow-up study, additional stipulations were imposed upon us.

Prof. Chen and Dr. D. Inbar reported on the *ad hoc* progress as well as on plans for the future, indicating that more concrete findings could be presented only upon obtaining the results of tests and questionnaires.

Y. Shohat, acting Director-General of the Ministry, inquired whether the researchers were aware of the objectives of their study - were they to justify the concept of Reform or merely to investigate curricula? (The transcript of the meeting was none too clear about this point; Mr. Shohat was apparently referring to the investigation of various operative decisions regarding implementation of the Reform).

Eliezer Shmueli, then Director of the Department for Educational System Development (today Director-General of the Ministry of Education) and Chairman of the Central Committee for Implementation of the Reform (appointed to this post on 12 June 1972 - several weeks before the meeting) noted that it was an oversight that no earlier meetings had been arranged between researchers and Ministry officials. He declared that:

There is an urgent necessity for checking out home room classes, degree of heterogeneity, etc. A detailed list of problems demanding investigation should be submitted in writing to the research team, so that by the next meeting we will be able to determine the points over which the Ministry has difficulties, in order to enlist the researchers' aid in solving these problems.

Prof. Lewy indicated that no single study could answer all questions and that other sources of information existed as well. Other participants offered comments, proposed programs and presented additional suggestions to the researchers. E. Peled summed up the discussion as follows:

- A. Objectives of research: The Minister demanded that current ongoing implementation of the Reform be accompanied by appropriate research, without waiting for its completion. Are our objectives indeed being achieved? If not, why not? How shall we correct matters? It is likely that research will indicate that only part of these goals have been attained; this provides us with an *ipso facto* definition of the study's objectives.
- B. We do not have time to wait 3-5 years for drawing conclusions; partial conclusions are useful as well. This is why it is important to have a steering committee which will direct research towards problems whose solution is feasible. This committee will coordinate all key factors within the Ministry, provide the researchers with material and assist them in carrying out their study.¹⁰

As a result of the above discussion, a research steering committee was appointed; five meetings were held between 24 October 1972 and 23 October 1974.*

The committee's first meeting, at which researchers were not present, was devoted to a discussion of its functions, which were ultimately defined as follows:

* The committee comprised senior Ministry officials who were involved with the Reform, namely: Chairman E. Shmueli - who was also Chairman of the Central Committee for Implementation of the Reform; U. Litvin - Director of the Ministry's Planning Department; K. Binyamini - Chief Psychological Adviser; Y. Friedman - Director, Mapping Department; Y. Shohat - Acting Deputy Director-General of the Ministry; and E. Shoval - Coordinator of the Pedagogical Secretariat.

1. To report research findings to the Central Committee for implementation of the Reform.
2. To provide feedback and guidance to the research teams.
3. To secure the cooperation of principals and supervisors with research staff.

Some committee members demanded active involvement in the research itself, through determination of policies and topics, examination of questionnaires, etc. Others maintained a more liberal approach, calling for involvement only in problematic issues. It was this latter approach which was eventually adopted by the committee chairman.¹¹ The first meeting with the researchers included a detailed discussion of the research.¹² It was also decided that the first interim report be submitted during mid-1973. Indeed, the first tentative research summaries were presented to Ministry officials during August and September of 1973, some two years after research had commenced.

Despite the key importance of this research project, the steering committee met to discuss the interim report a full six months after it had been submitted. No clearly-defined reason is evident for this delay, although three factors appear to have been responsible: a) firstly, the Yom Kippur War had disrupted the pace of daily life in Israel at that time; b) the data was only tentative in nature; and c) there were no outstanding findings which would allow conclusions to be drawn one way or another. Although little may be learned from the views exchanged at the steering committee meetings, it nonetheless appears that the overall atmosphere was one of considerable disappointment.

Most participants agreed that conclusions may be drawn from these findings only after additional work is undertaken, to be included in the interim report at a later stage (this referred chiefly to a comparison of background variables of junior high school pupils and elementary school pupils included in the sample) and only after a second interim report is submitted, including data on pupils who have completed grade eight.¹³

At that time, available data pointed to several findings about schools which had instituted the Reform, such as: an increase in integration, no significant difference in achievement (higher achievement for all pupils in classes with a low proportion of Oriental pupils), and poorer ability for self-evaluation of achievements by Oriental pupils as compared with the ability of students of European or American origins in this area.¹⁴

At the final meeting of the steering committee in its intra-Ministerial format, a number of members requested that the researchers publish findings of immediate significance with respect to policy-making, including such topics as guidance, slow classes and grouping. The researchers, however, could not meet this request and suggested that they be informed in advance of the subjects under discussion at the Central Committee for Implementation of the Reform, in order to present the relevant findings accordingly. The remainder of this meeting was devoted to a discussion of the final report's structure.¹⁵

Although the intra-Ministerial steering committee was not cancelled out completely, all matters concerned with research on junior high schools had gradually passed on to the Ministry's Office of the Chief Scientist and the professional steering committee appointed by it. A consideration of intra-ministerial treatment of research yields the following observations:

A. The Ministerial staff maintained a naive view regarding the steering committee's influence upon research. The steering committee could be influential primarily during the planning stages and when the final report was to be drawn up. Once the structure of research had been determined, it was most difficult to obtain mutual interaction on an active basis. This became clear to members of the steering committee each time they requested that the research include a topic which had not been agreed upon at the outset. The response was always the same - "Impossible".

B. The steering committee met only infrequently and its deliberations and decisions had no formal backing with respect to budgets and the scope of research - nor were they valid insofar as scientific criticism was concerned.

C. Most members of the steering committee were Ministry of Education officials and not scientists with academic credentials equivalent to those of the researchers. Hence their professional evaluations were not sufficiently valid or authoritative and could not answer the methodological questions which arose in the course of research. It appears that no serious attempt was in fact made at the solution of such problems.

THE INTERIM REPORT AND ITS RAMIFICATIONS

As indicated above, the first interim report was presented for the perusal of the Ministry of Education administrative staff during September 1973, some two years after the commencement of research. In their discussion and summary,

the researchers indicated the following reservations:

This preliminary report, which is based upon data relating to the entire sample (junior high school pupils and a control group of elementary school pupils), cannot serve as a basis for any conclusions relating to junior high schools; rather, it can serve solely as a basis for data processing which is to be continued further.

These reservations are grounded in three fundamental facts pertaining to the data available:

- A. The 84 junior high schools which were in operation in 1971/72, of which 18 were included in the research sample, are not representative of the educational system as a whole.
- B. The two seventh-grade groups sampled (junior high school and control) are not similar in nature, differing primarily in terms of the characteristics of their Oriental pupils.
- C. At least two measurements are required for deriving conclusions concerning change and progress; so far, data is available only for one measurement.

Despite their reservations, however, the researchers were able to determine that:

The overall results of achievement tests do not display a marked advantage for any of the groups studied (with the exception of the predictable advantage of non-Oriental pupils over Orientals in all groups). The reverse was in fact true - most consistently equivalent results for both groups were noted in achievement tests which are least dependent upon specific curricula (i.e. English and reading comprehension).¹⁶

In summing up, the interim report's findings hinted that reorganization of the Israeli educational system into elementary and junior high schools is more likely to have a wide-ranging effect socially and psychologically on pupils than upon their cognitive achievements. A more thoroughly researched and detailed response would be possible only following the accumulation, processing and interpretation of additional data.¹⁷

It is thus evident that the information presented to the Ministry after two years of research was meager indeed. This observation is not intended as a criticism of the research, as it had been planned as a longitudinal study; hence it was perhaps impossible to hope for more than what had been obtained. Nevertheless, there was an understandable feeling of disappoint-

ment among Ministry officials who read the report.

Four days after the interim report was considered by the steering committee, the topic of implementation of the Reform was discussed within the Ministry's Educational Executive - the highest-ranking non-political forum of the Ministry of Education. The meeting was opened by the Chairman of the Central Committee for Implementation of the Reform who said, among other things:

...

3. The reform was imposed upon us by a Knesset decision and we are not free to release ourselves from it - even in the light of budgetary limitations - except by means of another Knesset decision.
4. The Reform has coordinated and united the Ministry for joint activity. This also yielded the following fringe benefits:
 - a) a more balanced determination of school registration districts
 - b) review of curricula
 - c) it has dealt with social issues
 - d) concerted efforts have been made at constructing new schools and renovating existing ones.
5. The Reform is an efficient means of advancing disadvantaged pupils.
6. Integration activities may have to be toned down. Nevertheless, an Integration Law ought to be enacted within the current term of office, as progress will be impossible unless such a law is passed.
7. There was a certain slackening in attempts to solve the educational problems of the Reform.
8. Initial research findings will enable us to direct the educational and administrative aspects of the Reform accordingly.

H. Adler described the study and submitted particulars regarding the teachers involved. A. Lewy then referred indirectly to the Chairman's opening statement, claiming that: "Research does not actually serve to decide issues of policy and feasibility, as I see it. Rather, it complements data obtained from other sources, pointing out the relevant shortcomings and advantages." He then went on to discuss the interim findings regarding integration, study achievements and psycho-social issues. In addition, he indicated that "pupils in junior high school slow classes display poorer achievements than those who remained in elementary school."

The Director-General took the floor once the researchers had concluded their reports. He presented an extensive and detailed review of the Ministry's problems with respect to the Reform, including problems of a general social nature, as well as problems relating specifically to the junior high schools: their policies, educational staffs, professional and educational aspects, technological education, etc. The study had very little to say about these problems, especially in light of the objectives set for the research by the Ministry officials and the researchers themselves.¹⁸

Several participants at this meeting then expressed their opinions regarding the difficulties involved in the Reform, suggesting that the Ministry "take a deep breath" and concentrate on the educational aspects involved. Prof. C. Adler then took the floor; the minutes of that meeting indicate the following:

1. (Dr. Adler) objects to any let-up in implementation of the Reform, despite the financial difficulties, as this would not serve to cure the "anemia". (He later clarified that he was referring to the attitude of several senior Ministry officials towards the Reform).
2. The research sample comprises junior high schools with a high rate of integration, wherein the seventh-graders' achievements are no poorer than those of their counterparts in elementary schools where there is no such extreme integration. This is indeed an achievement.
3. Kibbutz and Yeshiva* high schools are thwarting attempts at Reform.
4. The Reform has improved teacher and curriculum standards.
5. Structural changes do not change content in and of themselves; however, no change in content has ever been achieved without a change in structure.
6. We should not give in to parents' objections, but rather overcome them through
 - a) the presentation of research findings
 - b) invitations to visit junior high schools
 - c) involvement of parents in the process of educational reform.
7. Transportation - weaker pupils are bused to better schools and not vice versa.

* Jewish parochial (mainly boarding) schools.

8. The problems of the slow classes are more serious than those of the grouped classes.
9. The second half of the school day which was to be devoted to extra-curricular activities and was to be an important component of integration, has been virtually unimplemented or put into effect only incidentally.¹⁹

Prof. Adler's statements - if the minutes are correct and recorded accurately - indicate a lack of differentiation between objective research findings and his personal views and opinions. This may have led others to accept his views as if they were in fact based upon research results. Some of the points raised by Prof. Adler had not been mentioned at all in the study; others, although considered therein, could not be utilized for drawing any final conclusions, as stated specifically in the interim report. At one point in the discussion, Prof. Adler stated that the research findings (which were then still indecisive and unclear) could indeed be utilized for convincing parents of the merits of integration. At another point, he declared definitively that the Reform had improved teacher and curriculum levels, even though final results were still pending. This was stated unambiguously, without his having indicated what his grounds or criteria were and without his having had the research data - as expressed in interim reports - before him at the time.

The deliberations continued and the Director-General appointed a committee to work out a suggestion for summarizing the discussion. Three months later, on 7 June 1975, the committee presented its recommendations. Although the study was not mentioned specifically, its influence may be discerned, to some extent, in the section dealing with slow classes.²⁰

The extent of faith in the ability of research to aid in the routine application of the Reform may be ascertained from the proceedings of a meeting of the Central Committee for Implementation of the Reform (29 September 1974), at which the Chairman of the Junior High School Committee declared: "We lack a continuous monitoring of administrative and pedagogical activities in junior high schools. The current academic research project, based upon a 12% sample, does not function as a follow-up study but rather investigates the achievements of the Reform". Another participant claimed that the absence of a follow-up study led to the raising of problems based upon intuition, rather than factual knowledge.

Once again, the topics covered by the research study - or at least those

which were touched upon in the interim report - were not the issues which interested senior Ministry officials insofar as implementation of the Reform was concerned.²¹

In this respect, it is interesting to note that the State Comptroller's Report included an investigation of the Ministry of Education's activities concerning the Reform. Without relating to the quality and professional level of this investigation, it is most useful nonetheless to note the topics considered therein:

Between February and July 1974, we conducted an intermittent investigation of the application of educational reform in Israel, some six years after its inception. We examined the pace of implementation, considered the resources invested in achievement of its objectives, and reviewed the status of junior high schools. This investigation was conducted among the administrative staff of the Ministry of Education and its central institutions, also including a sample investigation of fourteen schools at which the Reform had been instituted. This sample was determined according to various criteria, including the location of the junior high schools, the dates they commenced operation, their size and status in terms of connection with secondary schools...²²

The report mentions such well-known facts as the slow pace of implementation, preference for "easy" locations rather than those where implementation of the Reform would be most meaningful, linkage with academic secondary schools rather than comprehensive ones, the lack of suitable curricula, ignorance of learning achievements, etc. The State Comptroller's Report even considered research issues, such as the following:

In certain schools investigated by the State Comptroller's Office, it was impossible to discern any real progress from one level to another; in particular, no progress was noted from Group C, the lowest level, to Group B, the intermediate one.²³

The average drop-out rate among junior high school ninth-graders reached about 6% in the areas under consideration during the 1973/74 school year. This is a higher rate than the expected 5% postulated by planners of the educational reform. The drop-out rates for grades 7-9 at junior high schools during that year were greater than those indicated for parallel classes at schools where the Reform had not yet been instituted.²⁴

There were no clear diagnostic procedures for identifying prospective pupils for the slow classes. Under such conditions, pupils who belong in the slow class are liable to be included in a regular class (Group C) instead.²⁵

The report also indicates that the reverse may occur, with pupils who belong in Group C being placed in the slow class instead.

The State Comptroller's Report also dwelled upon the problems of district mapping, segregation within schools and improper operation of the latter half of the school day. The extensive consideration of these problems was based upon internal documents of the Ministry of Education, augmented by external testimony as well, such as data obtained from the Central Bureau of Statistics and the findings of field visits to the junior high schools themselves. In its summary, the report declared:

Institution of the educational reform demanded essential changes in many areas; thus, there was a need for coordinated activity among intra- and extra-Ministerial bodies and for substantial investments of professional, organizational and financial resources. Despite the limited budget and the difficulties in securing cooperation among certain involved parties, the Ministry succeeded in establishing 143 junior high schools throughout the country by the 1973/74 school year, improving the standards of teachers designated for these schools and instituting new curricula for a variety of subjects.

Despite these achievements, however, data available to the Ministry does not yet indicate that the Reform's goals have indeed been obtained. Our own findings display several shortcomings which are obstacles to achievement of these objectives: some junior high schools teachers are still inadequately trained; the preparation of new curricula and textbooks was not completed; some junior high school commenced operation in unsuitable buildings, while others began functioning without having all the necessary equipment at their disposal; difficulties arose regarding the organization of studies according to guidelines determined by the educational reform planners. The Ministry is aware of these shortcomings, several of which are also reflected in the interim report on implementation of the Reform, prepared by the Ministry during 1973/74. Budgetary limitations account for some of these difficulties, as well as for the relatively slow pace of the Reform's implementation. Nevertheless, among the findings of this report are a number of significant issues demanding action in which the financial factor does not carry decisive weight.²⁶

The findings of the State Comptroller's Report - in contrast to those of the interim research report - are outstandingly clear and comprehensive and relate directly to problems which concern the educational system. They indicate that certain matters may be investigated thoroughly and relatively swiftly. The State Comptroller's findings do not contradict those of the research study, with the exception of those relating to the subject of dropping-out. The Comptroller's Report mirrors reality - a reality which was essentially known by the Ministry administration as well. However, it does not express any causal connection among the respective phenomena, nor does it suggest any means of solution. Below, we shall consider whether the scientific study does succeed in doing so.

Ministry responses to the State Comptroller's comments were rather vague and of a general nature. Ministry officials apparently felt that the report yielded no new information and that reacting to its findings was a nuisance. Perhaps the report was somewhat embarrassing insofar as the public was concerned, but it appeared not to demand any additional consideration, nor to motivate new decisions or policies in any area whatsoever.

It would be interesting to determine whether the Ministry would have reacted in a similar manner to a report issued by well-respected university researchers. We will consider this issue as well in our evaluation of the Ministry's reaction to the final research report.

The 1974/5 Report on Implementation of the Reform includes several reactions to both the junior high school research project and the State Comptroller's remarks. The beginning of this report is devoted to an evaluation of the Reform's situation during its eighth year of implementation:

At the beginning of the eighth year of implementation of the Reform, which has reached some 40% of its projected application, we face the danger of losing the impetus necessary for completion of this key educational project and are threatened by perpetuation of a situation wherein two different educational systems exist side by side.

According to the author of this report, there are two reasons for this situation: financial limitations and "unprecedented, microscopic criticism to which the Reform has been subjected throughout." As nothing could be done about budgetary restrictions, the author felt that efforts should be concentrated on improving the Reform's image. This could be accomplished in one of three ways:

- 1) Emphasizing and publicizing the Reform's positive results:
 - The raising of standards for teachers
 - New curricula
 - New textbooks
 - The construction of school buildings
 - Progress in ensuring contact among pupils from varying social backgrounds.
- 2) Publishing the research report's initial findings in the following areas:
 - Teachers' attitudes towards the Reform
 - Parents' attitudes
 - The influence of junior high schools upon regional educational systems
 - Attendance rates of junior high school pupils.
- 3) Last but not least, directing public attention to the following facts:
 - The concept of Reform had still not been sufficiently finalized to allow for unambiguous evaluation.
 - The idea of "integration" was especially unclear and needed to be defined for areas with an ethnically homogeneous population as well.
 - The Reform had uncovered various problems of which we were unaware beforehand.

The report did relate to the initial research findings, after first stipulating that the final report would be ready only at the end of the 1975/6 school year. "Only on the basis of that report will it be possible for us to determine the extent to which the Reform has achieved its primary objectives, i.e. improving study achievement levels and promoting social integration among the different populations". The report then proceeded to indicate what appeared to be the preliminary positive findings: the attitudes of teachers and parents, the influence of junior high schools upon regional educational systems and the lower drop-out rate.

In contrast with this report's heavy reliance upon the interim report's findings - although no mention was made of the negative findings regarding counseling, the slow class, increased anxiety among Oriental pupils, etc. - its reaction to the State Comptroller's Report was defensive and apologetic. Firstly, it claimed that many shortcomings were brought about because of external factors, i.e. budgetary limitations and public opposition. Other faults were said to have been corrected since the Comptroller's comments

were issued. One important remark which had no solid background whatsoever was expressed: "The heterogeneous class succeeded in raising study achievements in the humanities..."²⁷

A meticulous analysis of this report reveals that the utilization of research findings was dictated by political realities and attitudes. The choice of findings was selective: there was an admixture of research results, evaluations and expectations, along with an unfounded methodological attack upon the uncomfortable findings indicated in the State Comptroller's Report.

LATER STAGES OF RESEARCH AND THEIR CONSIDERATION BY THE CHIEF SCIENTIST'S OFFICE

The Ministry of Education's Office of the Chief Scientist commenced more serious consideration of the research project toward the beginning of 1975. More than three years after the research had first begun, after all the data on pupils and teachers and most of the data about parents had been accumulated, the Ministry requested that the research be accompanied by a team of scientists. The first meeting of the research team with a number of senior university professors appointed by the Chief Scientist took place on 4 March 1975. The discussion focused upon various methodological problems. At the conclusion of this meeting, the Chief Scientist appointed Prof. Yehuda Amir to serve as chairman of the research steering committee. At first, the intention was that the committee would retain its original membership, albeit with a new chairman. However, Prof. Amir swiftly took upon himself the task of accompanying the research, with the assent of the Chief Scientist. One result of this meeting was the overall realization that the researchers could not possibly keep to their original timetable. Hence it was decided that the deadline for presentation of the final report be postponed for one year - i.e. to September 1976.

Until the time that the final report's first draft was submitted (May 1977), researchers frequently presented selected chapters for the perusal and comments of the Office of the Chief Scientist, the chairmen of the steering committee and the Central Committee for Implementation of the Reform and the Chief Inspector responsible for junior high schools. These excerpts were read thoroughly and aroused considerable comment, as a result of which changes were often instituted. There were also occasions when researchers requested permission from the Chief Scientist to publish partial research results in various scientific journals. Such requests were generally answered in the affirmative. Thus

it occurred that several research findings - particularly those dealing with dropping-out and with transition from elementary to junior high to high school - were published within the scientific community even before the final report was issued. One finding, however, was not released, namely the results of the tests of learning achievements. Although the first interim report had already implied that there was no significant difference in learning achievements between junior high school pupils and their elementary school counterparts, a fact which became even clearer in the final report, the Ministry and the researchers had avoided publicizing it. Once hints of this finding did appear in public, all responsible authorities were careful to state that no firm conclusions could be drawn prior to completion of the final processing.

As a result of increasing pressure from Ministry officials, interim research summaries were published in May and November of 1976. It is difficult to determine the extent of distribution of these reports within the Ministry, although it is clear that they did not reach the mass media. With the exception of the Chief Scientist, very few Ministry officials were involved with the research, although a memorandum prepared by the chairman of the Junior High School Committee for consideration by the Pedagogical Secretariat indicated a rather significant reliance upon its findings. Similar expectations were voiced by the Minister of Education in a newspaper interview, when he said:

At this time, thorough research on the achievements of junior high schools is underway. It has not yet been completed and will be presented to the Minister who succeeds me. This study will indicate both the problems and the achievements of the process of integration as carried out in junior high schools. I believe that no matter what the findings may be, it is not conceivable that the educational system will cease its constant struggle for integration of pupils from various ethnic backgrounds and strata. Integration is essential for social reasons, for overcoming prejudices and for promoting the true integration of all Jewish ethnic groups.²⁹

The Pedagogical Secretariat discussed the issue of junior high schools on 4 November 1976, with the participation of the researchers, Professor Amir and the Chief Scientist. The main topics discussed were the Reform's achievements in the area of learning achievements and its success in promoting integration. Ministry officials maintained essentially that the Reform was

not achieving its goals: two senior officials suggested linking junior high schools with elementary schools, while a third pointed out that the Reform had harmed vocational schools.²⁹ In the wake of this discussion, one of the participants of the meeting wrote a sharp letter, on 10 November 1976, to the high echelon Ministry officials:

...I am still under the influence of that very meeting. The two researchers, Profs. M. Chen and A. Lewy who are investigating the junior high schools and their achievements, have claimed that there is no difference between junior high school pupils and their counterparts in elementary schools insofar as learning achievements are concerned, despite the significant investment in additional hours at the junior high schools. With regard to more sensitive issues and social adjustment, it emerges that the situation is a bit more serious than when parallel situations arise in elementary schools.

The problems of integration were not solved by the Reform. The situation is perhaps most serious insofar as education in the religious sector is concerned, as selective preparatory classes for Yeshiva high schools were instituted in the Greater Tel Aviv region despite the objections of the Religious Education Division. The situation has not improved in regular State schools either, however. All experts maintain that the younger the pupils are when integration is first instituted, the greater its chances of success. Hence the problem must be dealt with primarily by the elementary schools.

In light of all this, I am amazed that not one person expressed openly that which many of us were thinking and even whispering furtively: Why? Why do we not have the courage to stand up and say: Cancel the Reform in education! Let us return to the 8 + 4 structure and utilize the buildings and resources thereby abandoned for training better teachers, setting up in-service programs and preparing improved study materials and more sophisticated study aids. I know that the decision to return to the previous structure will not be an easy one. There are serious factions which oppose me, but does their importance supersede the significance of the harm being done to our educational system under current conditions?³⁰

Despite all of the above-mentioned deliberations, the media's consideration of the Reform was not fed by research findings until mid-1977. The public attacks on the Reform were limited primarily to certain interest groups, such as the Israel Teachers' Union or those who opposed integration and were thus easily discredited. This lack of attention by the media is surely worth mentioning, considering that initial findings indicating no difference in learning achievements between junior high and elementary school pupils had reached the Ministry by September 1973 and taking into account the intense and vigorous disputes which had surrounded the Reform prior to its acceptance.

By late 1976, the researchers were apparently entering the final stages of preparing their report. Because of the topic's great importance to the public and in light of the methodological difficulties inherent therein and the allusions to findings on learning achievement, the chairman of the steer-

ing committee and the Chief Scientist decided to augment the steering committee with several renowned scientists, who would serve to provide a scientific seal of approval for the research findings or criticize them, if necessary. Dr. Leah Adar and Prof. O. Schild, two well-respected Israeli educational scientists, were selected for this purpose.

A draft of the final report was presented to the steering committee in mid-May 1977. For the first time in the history of the project, the event was given rather sensational publicity in the media. On 26 May 1977, a newspaper article in Ma'ariv was published bearing the following headline: "Findings of Two Secret Studies Presented to the Ministry of Education Reveal: Integration has not Led to any Significant Achievements; Educational Reform has Attained No Real Results." This article was written after the reporter's conversation with the American Prof. J.S. Coleman, who was visiting Israel at the time, in connection with a seminar held at Jerusalem's Van Leer Institute. However, only part of that newspaper article was actually based on this conversation; the rest included bits of information which the reporter had apparently gleaned from other sources. The article is replete with errors, misunderstandings and confusion regarding research findings. With respect to the study on the Reform, the only information presented by the reporter is to be found in the headline. Nevertheless, the article did arouse considerable public reaction; in its wake, the same reporter wrote a series of three more articles, published on July 11-13 of that year, also attacking the Reform rather severely. His criticism, which was published during a period of change in the government, fell upon sensitive ears indeed. The previous government was identified with the Reform, which was to be the "yeast" in its educational "dough". But how would the new government react? This question was all the more appropriate in light of the fact that the new Minister of Education, Zevulun Hammer, was among the leaders of the National Religious Party, which had been strongly opposed to Reform at one time; many of the party's supporters are in fact still opposed to institution of the Reform in State religious schools today. Such sensitivity to criticism influenced the press and senior Ministry officials alike, as well as the researchers themselves and the members of the steering committee.

Eliezer Shmueli, Director-General of the Ministry of Education and perhaps the senior Ministry official most identified with the Reform, responded to this series of articles, indicating his strong fear that the "socially valuable concept of integration" was eroding. He felt that publicity campaigns among parents would constitute an appropriate response to the criticism level-

ed at educational Reform: "I am aware that integration does cause problems, but the alternative is segregated education". Referring indirectly to research findings, Shmueli criticized the minority of parents who sought to escape the Reform under the guise of a variety of excuses - educational and otherwise. He then proceeded to state that even if the success of integration has not yet been proven conclusively from a scientific point of view, he believed that contact among the various groups was nonetheless preferable to the absence of the same. With regard to evaluation of the Reform, Dr. Shmueli stated:

It was only thanks to the Reform that the prestige schools at kibbutzim and moshavim and in cities and towns were opened wide to pupils who would never have gained admission to them otherwise. Were it not for integration, children from Haifa's Wadi Salib neighborhood would never have been able to study at the Reali School, nor children from South Tel Aviv at the Herzliya Gymnasia.

Dr. Shmueli went on to indicate that the integration process is not measurable since the very contacts among the students of different origins might by themselves form the basis for influences, the recording of which might appear in the more distant future.

Even if it is proven that the Reform did no harm and no good, I would still claim that it was of great significance because of its contribution in the affective/social sphere. If the State of Israel does really and truly desire to absorb immigrants from Eastern lands, then it must also pay the price, by standing up to the great social and educational pressures during the school years, even if this leads to a temporary slackening of the pace of studies.¹¹

The new Minister of Education had until then refrained from reacting to criticism in the media. He rather chose to await the Chief Scientist's evaluation of the study.

The researchers requested permission to react to the series of articles in a letter to the editor. After consultation with the steering committee, however, they decided to withhold their reactions until an official announcement of the research findings was published. Deliberations within the steering committee then went into high gear, with committee members investing over 100 hours of work in a highly detailed analysis of the findings, constantly bearing several important principles in mind:

- a) It was desirable to be as scientifically precise as possible, distinguishing between fact and opinion and striving for a balanced presentation of the findings.

- b) All committee members were convinced that the research findings raised no good reason to stop the Reform. They were likewise in accord concerning the essentially positive value and political and social significance of the integration process. Along with this, however, it was clear to all that an imprecise presentation of the findings was liable to serve as ammunition for those who opposed integration and the Reform. Committee members were most anxious over this possibility since they did not know the Minister's own views in this respect.
- c) It was vital that there be no repetition of the events which surrounded the publication of a report on the educational achievements of Israel's elementary schools (the Minkowitz Report), in which a disagreement ensued between the head of the research team and the Chief Scientist over the most socially significant findings, i.e. those relating to learning achievements. The researchers who had conducted the junior high school study desired the steering committee's seal of approval for their findings, while the Chief Scientist sought to avoid opposing the findings of a most important study for the second time in six months. This, he knew, would probably cast some doubt upon his authority as an official not beholden to the establishment, as well as upon the value of scientific research in general as a source of credible data for policy-makers.

According to the above principles, it is evident that both the researchers and the steering committee members sought to issue a statement which had been agreed upon by all parties concerned. If possible, it would be best to publish the report immediately, together with a statement by the steering committee corroborating the findings therein. However, the situation did not allow for such publication, as committee members' comments were too numerous and too important, relating both to the format of the report and the content of various chapters in it. The researchers and committee members ultimately agreed that publication of the report would be postponed for six months, during which time the report would be re-edited and several chapters rewritten completely. In the meantime, the committee would issue a statement - to be approved by all relevant parties - in the name of the Chief Scientist, detailing the main findings which could be reported at once. This announcement, issued at the beginning of the school year in September 1977, was rather compact in nature: one page of this four-page document described the research itself, while the remaining three outlined its findings. With regard to achievements, the announcement stated:

Research indicated that junior high school and elementary school pupils progressed in their studies at an identical pace, even though the former began, on the average, from a lower starting point. Researchers believe that it is indeed worth indicating that the progress of junior high schools pupils, under such conditions, proceeded at a pace no slower than that of their counterparts in elementary schools.³²

The words were chosen most cautiously: it was not said that the achievements themselves were similar, only that the pace of progress was the same, i.e. that the gap did not widen further. Nevertheless, the term "worth indicating" was interpreted both by the public and the Ministry to mean that there were relatively significant educational achievements within the junior high schools.

Regarding integration, the announcement declared:

Junior high school pupil populations are indeed more heterogeneous than those of elementary schools, not only in terms of ethnic and social heterogeneity but also with regard to learning abilities.³³

The announcement went on to indicate that junior high schools employed various systems of selection, some of which contravened the spirit of the Reform. It was also noted that fears of tension resulting from social contact had generally proved to be groundless. Other topics considered were the extension of the course of study, neighborly relations and attitudes of parents and teachers toward the Reform.

This concise document received extensive publicity in the press and in radio broadcasts. Despite the extreme caution exercised in its wording, the announcement was interpreted - against the wishes of the steering committee - as a report of positive findings alone relating to the Reform, meeting with criticism from various sources because of this. Both the Chief Scientist and the chairman of the steering committee later regretted that the announcement was issued at all, expressing their views publicly during a meeting of the Knesset's Education Committee.

Once the announcement was issued, the Minister began to reveal his own position on the Reform. His approach was most circumspect but still underlined clearly a lack of a strong commitment to the Reform in general and to the manner in which integration was then being implemented in particular. One of the first occasions on which the Minister discussed the Reform in public was during a newspaper interview:

Q: It has been said that you are considering repealing the Reform. Is this true? If so, wouldn't you say that it is a shame to have invested so much effort and money in it?

A: If money has been invested in something which is not worthwhile, it is indeed a shame, but that doesn't mean we should continue to do so - even if the structure erected is unsuitable - simply because money was spent. The Reform has been in effect for ten years and has succeeded in reaching 40% of the educational system. There are reasons for its not having extended further: manpower, physical, administrative and sectorial factors. Some IL 2 billion has already been invested and we must determine whether the results of integration have been positive or negative. We do not have any up-to-date research for the current year, which must be undertaken soon. The Reform was not instituted scientifically and thus does not lend itself to systematic investigation. Still, a study must nonetheless be carried out.³⁴

Prominent in the Minister's response was his desire to retain an objective point of view, without passing value judgements or committing himself in regard to integration and the Reform.

Against the background of this situation - considering the conflicting ideological positions, the change in government, the publication of controversial research findings and serious integration-related disruptions at the beginning of the school year - one may understand psychologically, if not politically, why A. Yadlin, the previous Minister of Education, took the initiative to make the following comment regarding integration:

Parents' opposition to integration has been increasing of late and there is a danger that the Ministry of Education may succumb to their pressure. Hence both the Ministry and the Knesset must reject categorically any attempt of the opponents of integration to base their arguments upon the research conclusions.³⁵

The Office of the Chief Scientist drafted a reply to be submitted by the Minister to the Knesset indicating the combined commitment to integration on the one hand and the increasing utilization of research on the other. The statement recalled several of the main research findings relating to integration, declaring that there was no justification whatsoever for its cessation.³⁶

In his response to Yadlin of 2 November 1977, however, the Minister chose not to mention the issues suggested by the Office of the Chief Scientist, stressing other aspects instead. After expressing his belief that parental involvement was a positive phenomenon, he continued:

I am keenly aware of the fact that many disputes were not directly concerned with the issue of integration itself, but rather with the places it was to be implemented and the manner in which it was to be effected - both of which are key factors for its success.

I must admit that decisions by the authorities regarding the location or concerning the relationships between "integrators" and "integratees" were not always the most appropriate or just solutions. Integration will not succeed if implemented arbitrarily; there are many factors which must be considered with caution and sensitivity, in order for the process to be efficacious in raising standards. The problem, as I see it - a problem which is of no less significance than the question of a drop in achievement level - is to determine the most effective means of promoting social inter-ethnic integration which will be a boon to all pupils.

One of the supreme goals of my Ministry is to raise the educational level and increase the intellectual awareness of the weaker sectors, of the disadvantaged pupils and of the Oriental Jews. To my mind, integration is an end and not simply a means.

Disadvantaged pupils often have marvelous talents but cannot discover, display, develop or apply them within the current educational system, which is more suitable for the middle and upper strata in terms of its methods, content and structure. The poor find it necessary to overcome deprivation, environmental difficulties and problems of a sociological and psychological nature. Poor pupils are subject to the limited capabilities and the vagaries of their parents and environment; they lack the appropriate supplementary tools at home and often attend schools which are themselves "disadvantaged". Such children must be rescued through precise diagnosis of their needs, preparation of special curricula and teaching methods and through establishing special programs for them at boarding schools and even local schools, providing a substitute for homes in which it is difficult to study. During the coming years, we will be devoting all our efforts toward this goal, so that such children will not drop out en masse and be exposed to the influence of the street, in order to help them matriculate and attend university or obtain high-level vocational training and in order to render them capable and suitable for their function as full and equal partners in society.

Regarding the study itself, the Minister said:

Studies available on integration do not offer a clear and decisive indication of results relating to knowledge and society. However, they do display certain encouraging signs, such as the decrease in dropping-out and the like. Along with this, however, it should be indicated that many such studies are based upon rather obsolete knowledge. Therefore I have asked the Ministry's Chief Scientist, Prof. S. Kugelmass, to prepare a series of up-to-date studies and investigations. In this manner, we will be presented with findings relating to the current situation and will be able to consider them at the opportune time.³⁷

An editorial which appeared in the newspaper Hazofeh, the Minister's own party newspaper, on 14 October 1977, stated the new Minister of Education's position in somewhat clearer language:

It was especially important to publish the findings (regarding the Reform) at this time, particularly in light of the numerous strikes held by parents who refused to send their children to junior high schools at the beginning of the term. These parents attempted to claim that so-called "research" determined that the Reform had failed. Now a study is published which indicates quite the reverse. Those circles which had anticipated negative results were keenly disappointed; some even promulgated the rumor that the Ministry had "doctored" research findings accordingly...

The public had received a mistaken impression regarding Minister Zevulun Hammer's attitudes towards educational Reform. The Minister is in fact an enthusiastic supporter of integration; he believes it to be one of the primary goals of the educational system and of Israeli society as a whole and intends to strive for its realization. However, in view of the problems which occasionally arise in the course of implementing integration, the Minister wishes to undertake a thorough investigation of its application, thus ensuring that the process will be carried out in a correct, efficacious and justified manner.³⁷

While Ministry officials generally considered the research findings to be a corroboration of the success - or at least the non-failure - of the Reform, Dr. Shalom Levin, Secretary of the Israel Teachers' Union, took an entirely different approach. In a series of newspaper interviews, Dr. Levin spoke his mind about the Reform and its attendant research:

I am doubtful about the Reform's success, particularly in light of the study undertaken. Junior high schools have excellent teachers with better academic training than those of elementary schools under the old formula. Junior high schools have received the very best educational equipment and benefit from the most advanced and efficacious pedagogical methods. In spite of all this, however, the achievements of junior high school pupils under the Reform do not surpass those of elementary school pupils under the old formula. This demands explanation.³⁸

I do not suggest that we get rid of the existing junior high schools, which now account for some 40% of 7-9th graders. The Reform's planners had promised the Knesset Committee that an investment of IL 150 million would be necessary to see the process through to completion. Instead, however, some IL 2.5 billion has been invested to date and the process is no more than 40% complete. Thus there is no reason to eliminate junior high schools or cut back in their standards. We accept graciously the decision by Minister Hammer to investigate the entire subject and to indicate in a serious and carefully-researched manner how the Reform has been of assistance, what it has contributed and which obstacles have thwarted achievement of its objectives. I claim that until the said investigation is completed and until we are certain of where the Reform has succeeded and where it has failed, continued implementation of the Reform should be frozen...

Two conflicting achievements - which could not be achieved simultaneously - were demanded of junior high schools from the outset, namely integration and the narrowing of gaps, together with progress in learning achievements. These goals are polar opposites: if integration is of primary importance, then the convoy can only progress at the pace of its slowest ship. If study achievements are of highest priority, then the best, most capable and most gifted pupils must be encouraged.³⁹

The public debate raged on, accompanied by sharp and sometimes violent conflicts as the Reform was implemented in various locations throughout Israel. At the same time, researchers and the Office of the Chief Scientist imposed a veil of silence upon themselves pending issuance of the final report.

At this juncture, the Minister decided to establish a public committee which would investigate implementation of the Reform. It is not clear precisely when and how he reached this decision, which was announced to the Chief Scientist at one of their meetings dealing with the research findings submitted. It is likely that the Minister's desire for such an investigation was not at all connected with the research project, in light of his essential attitudes towards the Reform. Nevertheless, there is no doubt that the research findings, which were rather unencouraging, indecisive and ambiguous with respect to key issues, did indeed constitute a motivating factor in his decision, which was likely to arouse negative public reaction. From the moment he announced his decision, the Minister virtually ceased to take an interest in the research. His activities relating to the Reform now focused upon appointment of the various committee members.

From that time on, any public statement made by the Minister relating to the research project has always been accompanied by two basic contentions: 1) the research is already obsolete, and 2) a committee has been appointed to investigate the entire subject of reform and integration. His decision to establish a public committee on the one hand, and the publication of abridged findings on the other, led to a situation wherein publication of the final report no longer had any serious political or social significance. It may be said that the final report was rather anti-climactic in light of the expectations expressed and the resources invested at the time the research commenced.

PUBLICATION OF THE FINAL REPORT

On 31 December 1977, the corrected drafts of the final report were submitted to the Office of the Chief Scientist and the steering committee for review. Committee deliberations again proceeded with great intensity and on 19 January 1978 its detailed comments were forwarded to the researchers. The final report was ultimately published in May 1978, bearing the title: "Junior High School Research - Processes and Results in Education: An Evaluation of the Contribution of Junior High Schools to the Educational System".

This report had 15 chapters and extended over 200 pages; it was thus considerably shorter than the first draft, which had 26 chapters and more than 500 pages. It represented a compromise between the steering committee's demands for a short and readable report and the researchers' desire to include most of the findings and conclusions arrived at thanks to their efforts. The researchers accepted most but not all of the steering committee's recommendations. Especially prominent was their refusal to moderate the tone of some of the report's concluding remarks.

The main research findings are well-known and have been mentioned frequently in this paper. It is therefore unnecessary to review them once again at this point. The public - especially journalists and media specialists - has apparently accepted the principal recommendations, namely "to continue with educational Reform and develop the junior high schools". Once the report was published, the Knesset's Education Committee promptly resumed its discussion of the findings indicated in it, devoting five sessions to consideration of the report and its implications, especially with regard to integration.

In opening the first such meeting, Member of the Knesset (and former Minister of Education) A. Yadlin repeated the factors which had motivated him to demand an urgent discussion of this topic:

I have a feeling - which I hope is unfounded - that we are hesitant about continuing to implement the educational reform. I deem it necessary to release the new Minister of Education and the Ministry administrative staff from such reluctance and to exert all possible parliamentary influence in order to ensure that implementation of the Reform continues.⁴¹

The main conclusion drawn by Mr. Yadlin from the research report was that selective methods were employed to an exaggerated extent in junior high schools. This, he claimed, sometimes cancelled out all effort: at integration.

The researchers then took the floor. Prof. Chen mentioned the intra-Ministerial steering committee, acknowledging the assistance rendered, but not concealing the differences of opinion: "One of the major bones of contention was the fact that our knowledge, at times, was not based upon data appearing in the report... The statistics and tables published constitute only a small part of the vast amount of data accumulated for each question, as summarized in only 4-5 pages."⁴²

The chairman of the steering committee, Prof. Amir, then responded, indic-

atings the committee's reservations toward such an approach. He outlined the committee's suggestion that researchers distinguish among three levels in their discussion of the findings:

The first level involves explanations or conclusions based upon non-controversial data, wherein the conclusions drawn by researchers are apparently the most correct ones. Level two relates to data for which various explanations and conclusions are possible; the researchers presented only a partial selection of these possibilities, while we suggested that the alternative ones be included as well. At the third level, we believed it legitimate for researchers to state, wherever necessary, that they were expressing their own opinions, without connection to the raw data under study.⁴³

Prof. Amir also mentioned the committee's view that findings relating to topics of selection - i.e. grouping and slow classes - did not justify conclusions at a level higher than the second one.

The ensuing discussions which took place during several sessions of the Knesset Education Committee provide us with a great deal of enlightening information concerning the problems of transmitting scientific messages - or those which are supposed to be based upon scientific findings - to political circles. We will consider this subject in greater detail below, indicating the issues expressed either overtly or covertly in the course of deliberations and which involve interaction between scientists and politicians.

The chairman opened the second session with several general comments about politicians' attitudes towards scientific knowledge:

I wish to repeat and explain my essential approach. In view of research results made available to me at the time we conducted discussions about the education reform, I had always claimed that integration alone is no panacea; the reverse may indeed be the case, as integration always entails complex educational and cultural problems. Nevertheless, integration in education is clearly a top-priority social necessity in Israeli society... I admit that I was most apprehensive that junior high school principals would avail themselves of the scientific conclusions and findings even prior to their publication. There was one finding in particular which would surely have saddened the late Mr. Zalman Aranne, namely: neither the junior high schools nor the Reform have realized their objectives with regard to junior high school pupils' learning achievements. To a certain extent, this is indeed a scientifically valid statement, but it was nevertheless possible to predict that the contradiction between establishing heterogeneous classes on the one hand, and striving to raise study levels on the other, would lead to the very tensions which plague junior high schools today.

...We must be prepared to pay a price for integration; the very essence of this research study indicates that we have not paid this price, but that study achievement results were reasonable nonetheless, as were those pertaining to social and emotional aspects. Another fact emerged in this aspect, however: if anyone did pay a price for implementation of the Reform, it was the weakest pupils.⁴⁴

We have quoted these matters at length for several reasons, primarily because it appears that an attempt was being made to rationalize research findings a posteriori. Mr. Yadlin had essentially expressed two thoughts which no one was prepared to voice when the Reform was first discussed:

- a) We should not expect study achievements to improve as a result of integration alone ("integration... is no panacea...")
- b) We must be prepared to pay a price in order to achieve the social objectives of integration (elsewhere, he indicated that he was referring to lowered study achievements).

It is also significant to note Mr. Yadlin's ambivalent attitude towards the study. On the one hand, he was apprehensive about its results, while on the other, he repeatedly relied upon it. His employment of such expressions as "To a certain extent, this is indeed a scientifically valid statement..." on the one hand, and "the very essence of this research study indicates..." on the other, represent clear-cut examples of this ambivalence, which is highly significant in our consideration of the influence of research upon policy-making.

Members of the Knesset S. Doron and M. Katsav spoke of the importance of integration; the latter, a young Oriental Jew and mayor of a development town, couched his words in especially trenchant and caustic terms:

I believe that our main objective is essentially the institution of total integration between the two pupil populations of varying development backgrounds, as each of the two is likely to enlighten the other. Hence it is not very important to me whether the study achievements have been significant or not, as this neither hinders nor helps the basic issue, which is the contact and integration of both groups of pupils. There is positive value in contact between disadvantaged pupils and those who have received better treatment; the former will gain more self-confidence, improve their powers of concentration, learn new concepts and sharpen their thought processes through joint games or even through discipline applied. Contact itself is an effective means for achievement of our key objective, namely the development of all kinds of latent talents and abilities.⁴⁵

One may note that MK Katsav was unimpressed with the opinions of research scientists and experts or by research reports which indicated that the mere mechanical application of integration did not suffice to improve achievements. Rather, he believed in the educational importance of contact itself.

MK Namir, well-known for her opposition to the Reform, sought to compliment the researchers:

I wish to congratulate the authors of this report; I feel they have displayed great courage in not concealing anything. I certainly believe that there is nothing which has been left unsaid in this report. Dr. Adler, whom I remember from our first discussions of the topic during the Seventh Knesset, was one of the greatest supporters of junior high schools. Nevertheless, he expressed some rather unflattering comments about the Reform, which was surely not very easy for him.⁴⁶

MK Namir's "compliment" displays an utter lack of sensitivity to the scientific principle which strongly emphasizes objective truth in research. At the following session, Prof. Adler referred to this very same point:

At the previous session, several MKs awarded us "high marks" for reporting precisely what we discovered, even though it appeared to them - justifiably, of course - that we had other expectations regarding study achievements. I wish to state that it is both natural and obvious that a scientist reports research findings as they are, without straying one iota from the truth. Anyone who does not do so is violating the principles of scientific behavior.⁴⁷

MK M. Elgrably, who had worked previously in the Ministry of Education's Planning Department, then declared that "all studies indicate that integration has a profound effect upon pupils".

MK Elgrably has an M.A. in economics and statistics and had devised one of the most important and efficient diagnostic tools for identifying disadvantaged children, the "Deprivation Index", which is based upon a number of socioeconomic background features. He then continued: "I believe that the best definition of a disadvantaged child is one who has not realized his potential..."⁴⁸

MK Elgrably's words - perhaps more than those of any other participant in the discussion - allow us to perceive how a change in function also changes one's individual perspective on various key problems. It was undoubtedly clear to MK Elgrably that his statements had no scientific basis whatsoever and that his new definition of the notion of "disadvantaged" has no operative significance and most likely no philosophical importance either. Nevertheless, he apparently made these statements because he believed doing so would be the best way to advance the interests which he believed to take precedence over scientific accuracy.

At the next meeting, in fact, Prof. Amir could not help but react to Mr. Elgrably's statements:

I wish to relate to Mr. Elgrably's remark that there is evidence that integration does improve achievements. According to my best judgement, there is no such evidence at all. Neither research findings nor theory have shown that integration must necessarily better the achievements of the upper groups.⁴⁹

While MK Elgrably did not state precisely what was attributed to him by Prof. Amir, such was undoubtedly the spirit of his remarks.

The Chief Scientist, Prof. Kugelmass, addressed the final meeting of the Knesset Committee and the research team, which took place on 7 June 1978. By this time, it was clear that the committee had concentrated primarily upon internal grouping in junior high schools, particularly with regard to the slow classes. Hence the Chief Scientist emphasized that the scientific proof for research findings in this area was not sufficiently powerful to warrant drastic changes such as cancellation of these classes. After indicating alternative explanations for these findings, he made one statement, toward the conclusion of his remarks, which proved confusing to many MKs:

However, the authors of the report base their analysis on things which are above and beyond the findings accumulated. They are experts. In the course of their research, they spoke with teachers, visited various locations, etc. We may rely on their expertise when they consider the slow classes unfavorably, but it should be clear that in this case, we are relying upon the researcher and not the research.⁵⁰

Several MKs asked the Chief Scientist about this incomprehensible distinction. When he attempted to explain once again the difference between relying upon the findings of a particular study and relying on the opinions of experts, MK Namir interrupted and asked: "If this is the case, why conduct a study at all?" To this, the Chief Scientist responded: "I do not believe we should expect too much from a study, as there are also differences of opinion among the researchers and scientists. I would say that the matter demands additional consideration."⁵¹

As to his own opinion of the slow classes the Chief Scientist continued: "I would say that we now have sufficient material which I would like to examine more thoroughly before making any concrete recommendation." At this point, MK Yadlin commented somewhat provocatively: "A clear answer: We need additional research before we reach any final conclusions".

Prof. Adler did not remain indifferent to these exchanges; his own comments related to the question of distinguishing between research data and researchers' expertise, as well as to what may be expected of research in general:

It is wrong to say that a researcher dealing with such a topic (study achievements) needs only to present a list of tables, while you politicians are to deal with them. We were asked to carry out this study not only as technicians but as men with experience, knowledge and comprehension of the educational system. It is certainly natural that after we

state the facts, we also indicate separately how we interpret them. We stand behind our interpretations and are in disagreement with the steering committee...

Therefore, I wish to state unequivocally that in our interpretation of the data before us - utilizing our experience and knowledge about teaching homogeneous and heterogeneous classes - we certainly see no reason for increasing homogeneity; in fact, we view this phenomenon as harmful. We do admit that the data alone do not indicate this unambiguously. However, if you expect research results to enable you to decide upon policies in an unambiguous manner, then I advise you not to allocate funds for research at all. To the best of my knowledge, no study has ever resulted in an unambiguous conclusion that, say, method X is the only correct one and all others are invalid. In general, we might say, for example, that 20% of the findings tend in one direction and 80% in the other.⁵²

Prof. Adler's comments are especially interesting in light of his admitting to the ambiguity of research findings, on the one hand and his clearly-defined public stand on the results of the educational reform, on the other. He even goes so far as to demand "calling the nation's educators to order" over education of the disadvantaged: "The fact that a person is disadvantaged is no reason for him not to study. Rather, it should motivate us to exert special efforts so that he does indeed achieve results."

The problem is that neither he nor other scientists are able to indicate how such efforts are to be exerted or - more precisely stated how one can work with heterogeneous classes and still ensure achievement and progress among all pupils.

The dispute among researchers, the Chief Scientist and the chairman of the steering committee was not yet over. Members of the Knesset Committee would now hear the summary of the steering committee chairman, which clarified the points of assent and contention once again.

The Knesset Committee's own prolonged discussions were summarized in its recommendations regarding integration. While its discussions had indeed focused upon the junior high school research report, its recommendations were of a general nature, dealing with the entire subject of integration. The Knesset Committee summed up its impressions of the study as follows:

The Knesset Education Committee became aware of the fact that under the educational Reform, more pupils of varying economic and social levels were integrated in junior high schools than in elementary schools under the previous formula. The Reform improved the staying power of pupils within the educational system and also succeeded in raising their study levels. The Reform was supported by parents; it improved neighborhood cohesiveness and raised the self-esteem of weaker pupils. Israeli children

do not attach any great importance to the ethnic origins of their companions. On the other hand, the Committee also realized that the very essence of change in educational structure did not raise the study achievement level of pupils, particularly the weaker ones, which is contrary to initial expectations. Nevertheless, the new curricula devised were instituted in elementary schools as well (i.e. those in which the Reform was not operative) and improved the study level of all Israeli pupils. Similarly, it emerged that teachers find difficulty in tackling the problems of heterogeneous classes. The educational system tends to circumvent this obstacle by setting up selective criteria for grouping slow and weak classes and by creating varying degrees of homogeneity in grade 9.⁵³

This summary is undoubtedly a reliable, concise and accurate reflection of the research findings. We will now examine the Knesset Committee's conclusions, based upon what it had discovered.

The Committee's recommendations regarding the organization of the system were:

1. To concentrate upon the implementation of integration in kindergartens and elementary schools and to remap registration districts.
2. To continue with implementation of the Reform, especially in places where it had already commenced.
3. To attach more junior high schools to comprehensive and vocational high schools.
4. To increase involvement of the kibbutz movement in solving the problems of disadvantaged children.

Regarding the organization of junior high schools in particular, the Committee determined that "the development of various means of selectivity in junior high school constitutes an obstacle to integration and hinders the progress of weaker pupils." It was therefore suggested that grouping be limited to two subjects only, that the slow classes be eliminated and that selection and guidance be applied primarily in grade 9, rather than grade 7.

Concerning learning achievements, the Committee demanded that the Ministry of Education devote more of its efforts to teacher training and the preparation of curricula for integrated classes.

With respect to research, the following recommendations were proposed:

1. The Committee notes with approval that so decisive a change as the Reform in the educational system was accompanied by close attendant research. Implementation of the Reform should continue to be accompanied by surveys and evaluative studies, in order to determine which factors ensure success at certain junior high schools while others are failing.

2. The Committee recommends that the 1978/79 junior high school study achievements be investigated, in order to discern the progress made since the last study, which was conducted in 1972/3.

The Committee also requested that the Ministry of Education declare its readiness to implement the above conclusions within one year.⁵⁴

It is clear that the research recommendations were accepted by the Knesset Committee as guiding principles, in some cases, this was true even when these recommendations were not endorsed by the Chief Scientist and the chairman of the steering committee.

Soon after these recommendations were presented, however, Mr. Yadlin was replaced as committee chairman by Mrs. Namir, a staunch opponent of integration who displayed no desire to pressure the Ministry into fulfilling the recommendations of her committee.

THE MINISTRY OF EDUCATION'S REACTION TO THE RESEARCH REPORT

As indicated earlier, senior officials of the Ministry of Education exhibited a highly restrained reaction to the study. About a month after the research findings were first published, the Chief Scientist issued his own views, in which he described the various difficulties encountered by the researchers in their evaluation of the Reform. These include both objective difficulties, which affect all research projects of such dimensions, as well as the handicaps which the researchers imposed upon themselves voluntarily in their desire to follow up on the previous research sample of Profs. Chen and Lewy. The Chief Scientist then related to the main findings in the areas of continuity, study achievements, social relations and parental attitudes; generally accepting the researchers' conclusions but expressing certain reservations, particularly with regard to certain instances of rather strong wording. The Chief Scientist's summary declared:

After reviewing the research findings, conclusions and recommendations, we are of the opinion that the suggestions presented in Chapter 15 result from four sources:

- A. Research findings and data
- B. Researchers' impressions upon visiting schools
- C. Knowledge available to researchers from the relevant professional literature
- D. Opinions and attitudes of researchers toward the various topics of their study.

These four sources constitute a legitimate basis for drawing up recommendations, bearing in mind that under such conditions, there is no one-to-one correspondence between such recommendations and the actual

research data or findings. Hence consideration of these recommendations necessarily deviates from the research/scientific plane and becomes more of a social or perhaps even ethical issue. In light of this, we will not deal with them herein. Those who deal with these recommendations in the future are advised to distinguish well between those which are anchored in research findings and those which stem from any other source.⁵⁵

There was full public coverage of the Chief Scientist's reactions insofar as transmission of the communique was concerned, although no significant public echoes were aroused.

At this juncture, the Public Committee for Investigation of the Reform Implementation was continuing its deliberations. This committee, appointed 20 December 1977, was charged with examining the implementation of the Reform, in fulfillment of the Knesset Committee's suggestion. A letter of clarification regarding the committee's appointment, dated 24 January 1978, states clearly that "the goal of the committee is to investigate the implementation of the Reform and not to determine whether or not here is any need for it. Similarly, the committee is requested to recommend to the Minister all improvements, corrections and changes which appear necessary as a result of this investigation."⁵⁶*

Regarding the issue at hand, as we are not fully aware of what transpired at the discussions of the said committee, it is important to note the written reactions of its members to the Reform research report, marking the similarities and/or differences between their recommendations and those proposed within the research report itself.⁵⁷

The committee's first recommendation "calls upon the Ministry of Education and upon local education departments to complete the Reform in the educational system and step up the pace of the implementation".

The research report's first recommendation likewise suggests "continuation of the Reform in the educational structure and in the development of junior high schools. It is especially significant to institute a uniform system in areas where the old and new systems are currently functioning simultaneously."

* This committee was headed by Prof. Amir, who was also chairman of the reform research steering committee and was well known for his professional involvement in the field of integration. Other members included Profs. Adler and Kugelmass, Dr. M. Shmida, three additional representatives of the Ministry of Education (Messrs. David, Ben-Eliyahu and Ron), two high school principals (H. Ish-Shalom and S. Shimoni) and two public representatives. Dr. Shalom Levin and Prof. A. Minkowitz, who had been appointed to the committee to represent the views of the Reform's opponents, resigned soon after its work commenced.

The public committee's report also alludes to this order of priorities (Recommendation 23: "First priority - areas in which the Reform was only partially implemented.").

The committee's second recommendation mentions the need for the application of wider ranging and more general mapping principles, echoing the research report's own second and third recommendations.

The third recommendation of the public committee deals with the principle of comprehensive secondary schools. While the research report does not present a parallel recommendation, its second suggestion, which is concerned with attaching junior high schools to secondary schools, does bear great similarity to the committee's views.

Committee recommendations 4-7 are concerned with the various aspects of devising tools to help teachers face the problems of heterogeneous classes. Recommendations 6-8 of the research report deal with the same subjects and suggest the same solutions.

The public committee's eighth recommendation demands that the Ministry of Education prevent junior high schools from carrying out certain activities which are accepted practice in secondary schools but are contrary to the goals of the Reform. "Their unique character will be expressed in various areas: class composition, activities for the prevention of dropping-out, advice and guidance, curricula, pedagogical methods, individual consideration of each pupil against the background of peer group problems, dealing with the problems of integration, class counselors, social education activities and in-service teacher training."

The research report's fifth and tenth recommendations express similar views; the former is concerned with selectivity in junior high schools and declares: "We recommend limiting the systems which function selectively in junior high schools, including violation of the district mapping principle, selecting pupils for homogeneous classes and groups and the creation of slow classes". The tenth research recommendation states: "We suggest continued emphasis upon the differences in atmosphere and educational objectives between the two systems, thereby ensuring the pedagogical and social uniqueness of the junior high school as a system suitable for adolescents."

The committee's ninth suggestion calls for the establishment of an efficient publicity campaign through "improvement of various tools for involving parents more closely in school objectives and securing their cooperation and

participation." The research report discusses the same topic in its 13th recommendation.

There are only two public committee recommendations which have no parallel in the research report: Recommendation 9, which suggests commencing integration at the elementary school level and Recommendation 10, which calls for appointing an administrative body to be responsible for progress in implementation of the Reform until its completion.

Generally speaking, it may be seen that the public committee accepted all the suggestions indicated in the research report, with the exception of the demand for an investigation of the creation of various educational climates. This is especially interesting in light of the fact that the committee and the researchers did not concur regarding the main research findings, i.e. that there is no difference between elementary and junior high schools with respect to study achievements. The committee report declared:

The junior high school study revealed that junior high school pupils displayed no better progress in study achievements than their elementary school counterparts (according to the results of examinations administered). Nevertheless, the committee was presented with various remarks and evaluations expressing reservations about this finding, coming from the Chief Scientist, the research steering committee and even from researchers themselves. In light of these comments, the committee's stipulation that the Reform has contributed positively to improving pupils' study achievements remains valid. Moreover, even the researchers themselves admit that test results constitute only one criterion for evaluation of study achievements and that the number of years of study must be taken into account as well. The committee accepts the view that a greater number of years of study generally expresses a rise in study level.⁵⁸

It is amazing that the Chief Scientist did not disassociate himself from this recommendation, which is surely not in line with his remarks of June 1978. In contrast, another member of the committee, Avraham Ron, was unable to concur with this stipulation. His minority opinion on this point of the committee's report declared:

Paragraph 1.2.1 (b) states that despite the findings of the junior high school research report - namely that there is no significant difference in study achievements between junior high school pupils and those in grades 7 and 8 of elementary schools - the committee's impression is otherwise. It appears to me that this impression does not carry sufficient weight in order to challenge research conclusions.

It should be remembered particularly that during the period of research, there were many more hours of study for junior high school pupils than for seventh and eighth graders in elementary schools. Nevertheless, no differences were found. I relate solely to study achievements and not to the "staying power" of junior high school pupils.⁵⁹

Had the Chief Scientist joined Mr. Ron in his views, they apparently would have carried far more decisive weight.

In summation, we note a marked similarity - at times even in the wording itself - between the recommendations of the junior high school research report and those of the Public Committee for Investigation of the Reform Implementation. This resemblance may result from a correct and accurate analysis of educational realities on the part of both bodies, or from the fact that four committee members were involved both in the research project and its steering committee.

First reports indicate that senior Ministry officials tend not to accept the public committee recommendations at face value. When the report was presented to the Minister of Education, top staff members expressed themselves in a highly non-committal manner. The Pedagogical Secretariat Chairman challenged the main recommendation, the Director-General regretted that the recommendations were presented at a time when funds were lacking for their implementation, while the Minister himself declared that he was not convinced by claims that the operation of two parallel educational systems was harmful. These are only recent developments, however; it is still too early to assess the extent to which the public committee's report and the Reform research report itself will influence Ministry of Education policies in the future.

SUMMARY

In describing the events which accompanied the Reform research, we attempted to supply the reader with as many details as possible. These were accompanied by comments intended solely to highlight those main trends and problems which appeared vital for the analysis and comprehension of the connections and relationships among researchers, research and policy-makers.

We have attempted to display the multi-faceted aspects of these relations, their ramifications in the media and the Knesset and the ambivalent attitudes toward the study evinced by the various bodies linked with it.

Notes

- ¹ Knesset Committee Report, p. 121.
- ² *Ibid.* p. 145.
- ³ Research proposal, p.5.
- ⁴ Interview with E. Peled, 29 March 1979.
- ⁵ Memorandum, meeting at Boyar School, 3 September 1971.
- ⁶ See Report No. 2 of Central Committee for Implementation of the Reform, meeting of 8 October 1972.
- ⁷ Letter to the editor by Minister Y. Allon, responding to critical article by Mr. Kesten in *Haaretz*, 18 September 1971.
- ⁸ "The Reform Research - Social Integration and Study Achievements in Junior High Schools". A letter to the Director-General of the Ministry of Education, E. Peled, signed by Profs. Cohen and Adler, 31 October 1971.
- ⁹ LEWY A. and CHEN M. Narrowing the Gap or Increasing the Deficit in Educational Achievements in the Elementary School. *Iyunim Beminhal Hahinukh*, 5736, 3-52.
- ¹⁰ Summary of meeting which took place on 4 July 1972, at which the junior high school research project was discussed.
- ¹¹ Memorandum, meeting of steering committee for junior high school research, 24 October 1972.
- ¹² Summary of second meeting of research steering committee, 17 November 1972.
- ¹³ Discussion of interim report on implementation of the Reform among Ministry of Education administrative officials, 24 February 1974.
- ¹⁴ See proceedings of this discussion.
- ¹⁵ Summary of fifth meeting of research steering committee, 23 October 1974.
- ¹⁶ CHEN M.; KFIR, D. and REGEV, M. Junior High School Research Interim Report No. 1 - Junior High Schools, Stage One: Achievements and Social Organization in Grade Seven. September 1973, p. 29 (mimeographed).
- ¹⁷ *Ibid.* p.30
- ¹⁸ Discussion of interim report on implementation of the Reform, *op. cit.*
- ¹⁹ *Ibid.* p. 6
- ²⁰ Recommendations presented at summary of discussion on interim Reform report, 6 June 1974.
- ²¹ Summary of 14th meeting of Central Committee for Implementation of the Reform, 29 September 1974.

- ²² State Comptroller's Report No. 25, p. 316
- ²³ *Ibid.*, p. 319.
- ²⁴ *Ibid.*, p. 320.
- ²⁵ *Ibid.*, p. 321.
- ²⁶ *Ibid.*, pp. 323-4.
- ²⁷ Report on Implementation of the Reform, 1974/5, 15 June 1975.
- ²⁸ Interview with Minister of Education A. Yadlin, *Maariv*, 21 June 1976.
- ²⁹ Proceedings of Pedagogical Secretariat Plenary Session 37/1, dealing with junior high schools, 4 November 1976.
- ³⁰ Letter to Minister of Education, Director-General of the Ministry and the Chairman of the Pedagogical Secretariat by Mr. A. Ron, Deputy Director of the Religious Education Division of the Ministry of Education, dated 16 November 1976.
- ³¹ Eliezer Shmueli, Director-General of the Ministry of Education, declares: "Parents are Attempting to Escape the Reform under the Guise of All Sorts of Studies". *Maariv*, 13 July 1977.
- ³² Excerpts from the Junior High School Research Findings. September 1976, p. 2. (mimeographed).
- ³³ *Ibid.*
- ³⁴ Young Correspondents Face the Minister of Education. *Maariv Lanoar*, 20 September 1977.
- ³⁵ Suggestion No. 200 for Knesset Agenda by MK A. Yadlin on Subject of Integration. Proceedings of the Knesset.
- ³⁶ Draft of Minister of Education's response to MK Yadlin's Knesset Agenda Suggestion on Integration.
- ³⁷ Response by Minister of Education to MK A. Yadlin's Suggestion to Knesset Agenda on Subject of Integration. Proceedings of the Knesset, 2 November 1977.
- ³⁸ Editorial in *Hanofeh*, 14 October 1977.
- ³⁹ Interview with S. Levin, Secretary of Israel Teachers' Union. *Hanofeh*, 14 October 1977.
- ⁴⁰ Interview with S. Levin, Secretary of Israel Teachers' Union, *Maariv*, 28 October 1977.
- ⁴¹ Proceedings of 64th Meeting of the Knesset Education Committee, 24 May 1978, p. 3.
- ⁴² *Ibid.*, p. 5.
- ⁴³ *Ibid.*, p.13. 102

- ⁴⁴ Proceedings of 65th Meeting of the Knesset Education Committee, 29 May 1978, p. 2.
- ⁴⁵ *Ibid.*, p. 5.
- ⁴⁶ *Ibid.*, p. 7.
- ⁴⁷ Proceedings of 67th Meeting of the Knesset Education Committee, 7 June 1978, p. 8.
- ⁴⁸ Proceedings of 65th Meeting, *op. cit.*, p. 11.
- ⁴⁹ Proceedings of 67th Meeting, *op. cit.*, p. 12.
- ⁵⁰ *Ibid.*, p. 3.
- ⁵¹ *Ibid.*, pp. 4-5.
- ⁵² *Ibid.*, pp. 8-9.
- ⁵³ Conclusions of the Knesset Education Committee regarding integration of pupils within the educational system; suggestion No. 200 to Knesset Agenda, *op. cit.*
- ⁵⁴ *Ibid.*, p. 23.
- ⁵⁵ Comments on Junior High School Research Report, Ministry of Education, Chief Scientist's Office, June 1968.
- ⁵⁶ Letter of appointment, Public Committee for Investigation of Reform Implementation, 20 December 1977.
- ⁵⁷ This statement is based upon comparison of the recommendations indicated on pp. 192-6 of the junior high school research report with those of the public committee, to be found on p. 3 of their own report.
- ⁵⁸ Public Committee Report, p. 11.
- ⁵⁹ *Ibid.*, p. 69.

EVALUATION OF A NATIONAL MATRICULATION EXAMINATIONS SYSTEM

G. Ben-Shakhar and B. Nevo

A BRIEF OUTLINE

While most evaluation studies in education deal with curriculum development projects, this article focuses on the question of whether a testing program can also be assessed via the standard tools of evaluation methodology. The Israeli high school Matriculation Examinations (ME) project is the "target" of this theoretical discussion.

A short description of the ME is provided, and the positive and negative effects hypothesized to be associated with the ME are specified. In the next section, a proposed design for an extensive series of evaluation studies of the ME is outlined. In this somewhat complex design, a special effort is made to deal with as many aspects of the ME as possible.

Limitations of the proposed study are presented and discussed in the last section of the article. The question of combining several outcomes (positive and negative), which are qualitatively different from each other, is raised. The decision as to whether such an evaluation study is feasible and promising enough to be carried out, is left open.

1. DESCRIPTION OF THE MATRICULATION EXAMINATIONS (ME) SYSTEM IN ISRAEL

1. The Requirements for a Matriculation Diploma.

Every high school graduate in Israel must take external examinations in a number of courses in order to qualify for a matriculation diploma. These examinations are external in that their construction, administration, and grading are executed by the Ministry of Education and not by the schools.

The Israeli high school program is based on a unit system with courses being offered at various levels. The difference between the levels of a

single course is defined in terms of the number of units a student has to take. Each unit requires three weekly hours throughout one school year, or one weekly hour throughout each of the three years of high school. For example, a student who takes a course at a five unit level must enroll in this course for each of the three years of high school, five hours per week (the normal senior high school program lasts three years). The levels offered usually range between two to five units. The Ministry constructs a different examination for each course level.

Students are required to take examinations in a minimum of twenty units. At least two courses must be taken at the four or five unit level. Certain courses (e.g. Hebrew language, literature, and one foreign language) are obligatory for all students, while other courses are optional.

Some examinations can be taken at the end of the tenth or eleventh grades, but most are taken at the end of the last year of high school.

2. The Nature of the ME.

The examinations are constructed by special committees. For each course the committee is nominated by the Ministry supervisor responsible for that subject. The committee usually consists of the supervisor as chairman, a senior teacher and one or two university professors.

The examination usually consists of essay questions, though in a few selected courses (e.g., English, Biology) part of the examination consists of multiple choice questions.

3. Administration of the ME.

Each examination is administered at a specific time, simultaneously throughout the whole country. The students take the examinations in their own schools. All ME are administered during a single month in the summer vacation. Students may take examinations again if they fail or if they wish to improve their grades.

4. Grading of the ME.

Every examination is graded by two experienced teachers independently. In most cases this work is carried out in special assessment center, where all the graders are assembled and supervised. The examination score of each student is determined by averaging the two scores given to him by the independent graders. The final matriculation grade for each course is the

mean of this grade and the grade given by the teacher of the course on the basis of the student's performance throughout his or her high school years. Only the final matriculation grade appears on the matriculation diploma.

5. *The Volume and Cost of the Matriculation System.*

Approximately 20,000 students take the ME every year; each student taking between five and eight different examinations. The Ministry constructs over five hundred different examinations per year, and the cost of the whole operation is approximately four million American dollars.

II. EDUCATIONAL GOALS SUPPOSED TO BE ACHIEVED BY THE ME

The Ministry of Education in Israel has never officially specified the goals which it expects to achieve through the implementation of the ME in high schools.

There is nevertheless no doubt that such goals, or desired functions of the ME, do exist for some people, or otherwise the program would not have continued for so many years.

In order to investigate what the desired goals are, two sources were used: 1) a survey of the literature (Dockrell, 1980; Nevo, 1980), and 2) interviews with several dozen educators, administrators, school principals, teachers, students, parents and researchers.

Almost every writer and interviewee was critical of certain aspects of the program in its present form, but most were in favor of maintaining some kind of ME. Those who were in favor of the ME seemed to agree about the basic, broadly defined, positive effects to be gained by the ME. The following list reflects the consensus of those in favor of the ME:

1. The matriculation grades an individual student receives provide him with information regarding his achievements in the different courses. Among other things, he can use this information in making decisions regarding his future career.
2. Teachers, principals and national authorities can use the data from the examinations as an objective feedback system. If the necessary statistics are properly extracted and handled, problematic classes or schools can be identified, new teaching methods can be tested, and deficiencies in the national high school curriculum can be detected.

3. ME grades can be employed by universities as a selection device for the purpose of screening candidates for the various undergraduate programs. Since a matriculation diploma is a prerequisite for registration, it is the student's attainment at the ME that counts. Universities also benefit from the homogeneity of the knowledge which their new students bring with them (see point 5 below); thus they do not have to invest too much in preparatory courses.
4. External examinations, such as the ME, create special and extra motivation among high school students. The ME encourages the students to learn more thoroughly and intensively than they would do otherwise. Some educators consider this source of motivation as the only available one, especially when the number of students in the class often is higher than thirty five. According to this view, the ME is the best motivational strategy toward higher standards of achievement.
5. In order to prevent their students from failing in the ME, high schools have to follow very closely the curricula of the courses as dictated by the Ministry of Education. The Ministry openly uses the ME as a curriculum-imposing device. The Ministry officials hope that this centralization helps in maintaining homogeneous standards of achievement all over the country.

III. EDUCATIONAL DAMAGES OF WHICH THE ME ARE ACCUSED

Other Israeli scholars oppose the idea of the ME and would like to see it eliminated. First, they disagree with the opinions quoted in the previous section. Second, they point to various negative effects which, in their opinions, are direct results of the testing program. The following is an attempt to summarize these arguments against the ME:

1. The examination induces high tension and very strong anxiety among many of the youngsters. In certain individuals this may turn out to be a traumatic experience with enduring effects.
2. The ME presents a constant (or chronic) source of conflict between high school students and their parents on the one hand, and the Ministry of Education on the other. Every year hundreds of letters of complaint are sent to the Ministry or to the media. The attacks run from accusations regarding the ideology of education to technical flaws that are found from time to time in the ME procedures.

could serve to evaluate several goals simultaneously. We would like to suggest that the best design is that of a randomized experiment (Boruch and Rindskoff, 1977). The sampling units in this case must be classes (or schools). It is therefore necessary to draw a random sample of $(2N)$ classes representing the different geographical areas of the country and the different types of schools. Half of these classes (N) will constitute the experimental group for our proposed study, and the other half will serve as controls.

Students in the first group will not take the ME; they will only be graded by their teachers. These students will have to be guaranteed, of course, that they will not be hurt in their future careers by the lack of a standard matriculation diploma. This last condition is important from an ethical point of view as well as from a methodological one. Any long-term follow-up of the future careers of the students is based on the assumption that the mere fact that these students did not possess a matriculation diploma will not at all interfere with their careers in any way. The other N classes will constitute the control group. Classes in this group will continue with routine participation in the ME program. Since classes will be randomly assigned to the experimental and control samples, we can assume that they will be similar to each other in all respects except for the requirement of the ME. This procedure should be repeated for six to eight years.

A feedback system based on ME grades should be constructed for all classes in the control group. Feedback to each teacher should include detailed data of the performance of his/her class including components of the ME (individual questions, chapters, etc.). National norms should be provided along with information about the particular class to enable meaningful interpretation of the class grades. Norms should be based on the performance of all students who took the specific examination in a given year.

In order to measure effects, feedback will have to be provided to the teacher for several years. Hopefully, this will also result in better utilization of the teacher's time and efforts, and ultimately bring about better performance by his or her students in future years. In order to test this hypothesis, standardized tests (which will provide the alternative to the ME) will be designed each year in each course. Test construction will be on the basis of the declared curriculum, independent of the ME. These tests will be administered to both the control and experimental groups each year, and the data analyzed on a year-by-year basis to examine possible improve-

ments in the performance of the experimental classes after receiving the feedback information. In other words, it seems to us that in order to evaluate the effects of feedback, we must provide the teacher with an opportunity to learn from the feedback, and one should not expect immediate results in terms of student achievement, but rather improvement in teaching methods that would influence achievement in future years. This line of reasoning leads us to a time-series type of design (Box and Jenkins, 1970) in which the focus of research is the teacher. The hypothesis that feedback has a positive effect will be tested by differences in the experimental and control time-series.

Another type of feedback mentioned in Section II is the feedback to the individual student. An attempt to evaluate this type of effect of the ME also requires a longitudinal study. (It should be remembered that most high school graduates in Israel serve two to three years in the army immediately after graduation, and thus postpone their professional career). It is doubtful whether it is at all possible to isolate and assess the role of ME feedback in career choice decisions so many years after graduation. However, if attempted, evaluation of the effect of this type of feedback could be based on the above-mentioned design. Two possibilities present themselves in this type of a study: 1) to study the decision process in career choice, trying to assess the weight of ME results in these decisions; 2) to study the outcome of career decisions, and to compare the control and experimental groups on dimensions such as job satisfaction and success.

Another goal of the ME mentioned in Section II is their motivational aspect. Again, the motivational effects could be studied using our experimental approach, with the aid of proper standardized tests. In fact, this could be done using the tests as mentioned earlier. This time the analysis should aim at a comparison between experimental and control groups, rather than comparisons over time.

Another possible advantage of the ME is that students have homogeneous knowledge when starting their higher education. This aspect of the ME could also be studied using our experimental design. The control and experimental groups could be compared on a number of standardized tests just before the beginning of their higher education. At this point we could compare the two groups for homogeneity and for the level of prior knowledge necessary for academic studies.

A very important function of the ME from the point of view of the Ministry of Education is its role in imposing a standardized curriculum. Our experimental design could be used to compare the content being taught in the experimental and the control groups. This could be achieved by the use of observers who would visit the different classes. It is, of course, necessary that these observers be "blind" regarding the identity of the classes. The goal of these observations is to look for possible differences in the way different teachers follow the curricula of the courses as dictated by the Ministry. Negative effects of the ME (see Section III) could also be evaluated by means of the above-mentioned design. Experimental and control groups could be compared on anxiety level, attitudes towards the educational system, frequency of misconduct and dishonest behavior, creativity, initiation and involvement among students and teachers during high school, etc.

An important aspect of the ME is their role in selecting students to the universities. One obvious question regarding this aspect relates to the validity of ME grades for predicting academic success. Since most institutions use the ME results along with other predictors (such as aptitude tests), the validity question should relate to the marginal contribution of the ME to the general predictive value of the combination of all the other variables used. An even more subtle formulation of the validity question could relate to the alternative of the ME, namely, the grades given to the students by the class teachers. The contribution to predictive validity of the ME grades should be compared to the possible contribution of grades given by the class teachers. It is possible that the grades might turn out to be a more efficient predictor of academic achievement than the ME.

V. LIMITATIONS OF THE PROPOSED EVALUATION STUDY

We have proposed a single study to evaluate several positive and negative aspects of the ME. This study is based on a randomized design. The data to be collected is heterogeneous and composed of observations in classes, attitudinal questionnaires, and different achievement tests to be administered at the termination of high school and again just before the beginning of the students' higher education. In addition, questionnaires should be administered several years after graduation to clarify the role of the ME in career planning.

It is clear that although our study was designed as a randomized experiment, the conditions in the field could interfere with this randomization, mainly because of attrition of both students and teachers. We should, therefore, regard the design as a quasi-experimental one, which somewhat reduces the validity of the conclusions. To compensate for the loss of validity we propose a time-series design to study the effects of feedback on teachers. If the ME is effective, we should expect a gradual improvement in achievements of the control classes as a function of feedback.

A major methodological problem with our proposed study is that to a great extent we rely upon achievement tests. One could justifiably make a case for the view that the tests used as a research tool are in fact alternative to the ME and, in fact, there is no real "no testing" experimental group in our study. In addition, there could be some effects due to repeated testing (the control group undertaking the ME and our battery of achievement tests). A possible solution could be that the standardized tests be administered shortly before the ME, and with no prior announcement. This might be technically difficult to carry out, but it seems workable.

The fact that we deal simultaneously with several research questions in a single study, complex and sophisticated as it is, may be convenient, but it creates methodological problems. Specifically, it would be difficult to isolate effects; any differences in achievement between the control and experimental groups could be attributed to both the feedback aspects and the motivational aspects of the ME. We mentioned earlier that feedback information should affect achievement only gradually and after several years, while motivation-arousal is expected to have immediate results. But this may be too small a difference to be detected. To solve this problem, one could create an additional group of classes that will take the ME with no feedback information system.

Another methodological problem relates to the generalizability of the suggested study. Factors such as the Hawthorne effect and the John Henry effect could lead to results that are specific to the experimental setting.

Some of the goals of the ME are long term by nature. We suspect that those goals will be especially difficult to evaluate. It will be technically difficult to identify and question students many years after high school graduation. Even if enough students can be located, it would still be hard, if not impossible, to assess the role of the ME in professional and career decisions.

It should be noted that not all possible aspects of the ME can be measured by the proposed model. For instance, social factors such as equal opportunity, avoidance of favoritism, etc., are considered as positive aspects of a national ME system. The quantitative measurement of these factors is very complicated and beyond the scope of the present paper. However, even if it is assumed that every single effect could be evaluated, the question still remains as to how these effects should be combined, i.e., what the relative weights of all the aspects involved in the ME project are, and whether these can be translated into practical decisions (Edwards, Guttentag & Snapper, 1975). At any rate, in our view, the proposed design is the only approach for an empirical evaluation of the ME. Therefore, the crucial question here is one of policy: should Israel neglect the whole idea of evaluating the ME project (because it is a complex, risky and difficult task to carry out) and continue spending four million dollars every year, or should it start an evaluation study, with certain limitations as described above, hoping that it will lead, in the long run, to rational decisions?

REFERENCES

- BORUCH, R.F. and RINDSKOFF, D. On Randomized Experiments, Approximations to Experiments, and Data Analysis. In: L. Rutman (Ed.) *Evaluation Research Methods: A Basic Guide*. Beverly Hills, California: Sage, 1977, 141-176.
- BOX, G.E.P. and JENKINS, G.M. *Time series analysis: Forecasting and control*. San Francisco: Holden Day, 1970.
- DOCKRELL, W.B. (Ed.) *The Impact of Tests on Education*. Princeton, N.J.: International Association for the Educational Assessment, 1980.
- EDWARDS, W., GUTTENTAG, M. and SNAPPER, K. A decision-theoretic approach to evaluation research. In: I.E. Struening and M. Guttentag (Eds.) *Handbook of Evaluation Research*, Vol. I, Beverly Hills, California: Sage, 1975.
- NEVO, B. (Ed.) *Matriculation Examinations in Israel*. Tel Aviv: Am Oved, 1980.

SHARING RESPONSIBILITIES FOR EVALUATION: COOPERATION BETWEEN THE DEVELOPMENT AND EVALUATION TEAMS

Arieh Lowy

Increased demands for systematic evaluation of educational processes of various types have led to the establishment of Evaluation Units operating within the framework of administrative bodies and development centers. These units are asked to continuously produce feedback and evaluative data summaries relevant to the needs of the organization within whose framework they operate. These Evaluation Units have usually specialized in evaluation studies of a specific type. In many cases they have had to carry out several evaluation studies simultaneously, each of these being shaped according to a common pattern or design. In other cases they have had to carry out similar studies recurrently or repeatedly, i.e. immediately after finishing a particular study or a particular series of studies, they have had to start a new one, following exactly the same pattern and design as the previous one. The practice of establishing Evaluation Units within the framework of institutes constitutes a new development in the field of evaluation, which has implications both for the management of evaluative studies and for the conceptual definition of such studies.

In the past, evaluation studies were usually carried out by "free-lance" enterprises on the basis of *ad hoc* contracts, and thus even if a lasting relationship evolved between the evaluation team and the team whose work was evaluated, it was a very limited one. In contrast, Evaluation Units operating in institutional settings are exposed to ongoing contacts and there is intensive interaction with the teams whose work they are supposed to evaluate. They are subject to the control of the same authorities who assume responsibility for the target of evaluation and participate in many activities together with other units of the same institute.

Evaluation Units entrusted with carrying out recurrent evaluation studies are now found in many diverse settings. For illustrative purposes, two

examples will be mentioned here. The ideas presented in this paper reflect experiences accumulated within the framework of one of these examples. The striking similarities in the operational mode of a great number of Evaluation Units nevertheless justifies close examination of their working practices, and suggests that a Case Study type description of one Evaluation Unit operating in a particular institutional setting may serve as the starting point for making some generalizations about Evaluation Units in general. First, some details of the two examples will be described.

The first example is that of Evaluation Units operating within universities. Several universities maintain units which are supposed to provide evaluative information about the activities of their academic departments. At some universities, the activity of such a unit is restricted to evaluating the teaching quality of the faculty members; at others the Evaluation Units adopt a more comprehensive approach and examine a broad range of characteristics of the academic units, such as the style of leadership, cooperation between faculty members, services rendered to the community, impact on the community of scientists in a certain discipline, etc. No matter how the target of evaluation is defined and what the evaluative criteria are considered to be, the Evaluation Units are expected to carry out studies on all the academic units according to a predetermined conception, and are supposed to produce a series of evaluation reports. Thus for example, the Evaluation Unit of one large university is requested to prepare a report on each of the 100 academic units operating within the university during a cycle of seven years, i.e. each year the Unit has to prepare approximately 15 such reports. While the evaluation report of one academic unit may be based on data of a different type than that of another unit, and the reports may consequently be quite different from each other, if one examines the whole series of 100 reports one will find that most of them are shaped according to two or three basic patterns. Given the similarity of the study patterns, the term Serial Evaluation Activities (SEA) will be used for describing activities of this type. Serial Evaluation Activities encompass a single coherent set of studies consisting of a series of replicated activities. It should be noted, however, that while Serial Evaluation Activities can only be performed by a body which operates on a permanent basis, they need not necessarily be done within the framework of an institute which assumes responsibility both for the evaluation activity and for the target of evaluation. Serial Evaluation Activities can be performed by institutes whose sole responsibility is carrying out the evaluation studies. Thus, for

example, the National Assessment of Educational Progress is a series of studies of educational achievements in various subject fields carried out in cycles of ten years by a permanent institute whose sole responsibility is planning and carrying them out. This institute has established a pattern of collecting data, reporting results, etc., and the work done by it fits the above given definition of Serial Evaluation Activities (SEA). The focus of this paper, though, will be on Serial Evaluation Activities which are carried out by Evaluation Units operating within the framework of institutes assuming responsibility both for the evaluation itself and for the target of the evaluation.

This leads us to the description of the second example of an Evaluation Unit operating within an institute.

The Ministry of Education and Culture of Israel established the Israel Curriculum Center (ICC) in 1966, and this Center assumed responsibility for producing new curricula and innovative instructional materials for the whole educational system. Approximately thirty different curriculum projects were undertaken within this Center. A separate team assumed responsibility for each individual curriculum project. Thus, thirty teams operated within the Center, each assuming responsibility for developing and implementing a single curriculum project. The Evaluation Unit established within the framework of the Israel Curriculum Center assumed the responsibility of conducting evaluation studies related to each of these projects. Due to the similarity of tasks *vis-à-vis* each of the different projects, the Evaluation Unit carried out studies of the type which were referred to previously as Serial Evaluation Activities (SEA).

The subsequent sections of this paper will describe some aspects of the evaluative work carried out by the Evaluation Unit of the Israel Curriculum Center. It will deal with the following issues: a) the organizational structure of the Evaluation Unit; b) a conceptual framework for evaluative studies; c) working strategies; and d) the clientele. Using colloquial terms, one may refer to these issues as who does what, how, and for whom.

It should again be stated that the purpose of this paper is to call attention to some general features of Serial Evaluation Activities conducted within the framework of institutes which are responsible for both the evaluation and the target of the evaluation. Accordingly, the detailed description of the operation at a single institute is given here with the aim of deriving some generalizations, rather than for the sake of presenting a historical sketch of a unique occurrence.

THE ORGANIZATIONAL STRUCTURE OF THE EVALUATION UNIT

The formal structure of institutes may be an outgrowth of historical developments, but, in the case of new institutes, it usually reflects a certain conception of roles and of hierarchical role division. This does not mean that the formal structure fully describes the relative importance of each unit within an institute. It is well known that units having equal formal status within an institute frequently differ from each other with regard to their actual power, their access to resources, and their impact on decision-making processes. Nevertheless, given the fact that the Israel Curriculum Center is a relatively new institute, it is of interest to study the formal status of the Evaluation Unit within its framework. Such a study, even if it does not reveal all the relevant facts about the actual status of the Unit within the Center, nevertheless may reveal the conception of those who established the Center as to the role evaluation should play within its framework.

Since the Curriculum Center was set up with the aim of introducing a curriculum reform in the school system, its major domain of activity has been the production of innovative educational programs and instructional materials. Evaluation was meant to aid the process of production and was not considered as an aim in itself. Its purpose was to serve the needs of those who prepared the new curricula. This conception of role division and role hierarchy was reflected in the formal structure of the institute. The major units operating within the institute were the Curriculum Projects. A Curriculum Project was defined as the production of a new program (including textbooks, teacher's guide, instructional aids, etc.) in a single subject area for a group of classes in schools of a certain type. Thus, for example, Biology for the Disadvantaged Learners in the Middle School, or Bible for the Upper Section of the Religious High School, were defined as Curriculum Projects. Each project constituted an organizational sub-unit within the Center, having its own team and its own project director. In addition to these production units, several service units were established in the Center, such as the Graphics and Printing Unit, the Library and the Evaluation Unit. The Evaluation Unit, from the very outset, has been considered as a service unit, and was supposed to serve all production units. The fact that the Evaluation Unit was similar from the point of view of its formal status and staff size to the majority of the thirty ongoing Curriculum Projects, determined the parameters of its activity. Due respect was given to evaluation

within the framework of the Center. While to a certain extent autonomy was granted to all Curriculum Project teams, at the same time the formal organizational structure manifestly conveyed the message that by no means should the Evaluation Unit have claims to privileges which were not granted to other Units of the Center. The formal structure reflected the manifest goals of the Center, but at the same time also gave expression to the basic philosophy underlying the operation of the Center, according to which empirical support should be sought for improving the quality of the new programs and for increasing their success in the school system.

By establishing an Evaluation Unit the stage was set for starting the evaluation activities. It was clearly evident that a small evaluation team could not provide all necessary services for all Curriculum Project teams operating within the Center. It was important to find a way which could ensure the conducting of evaluation studies for all the Curriculum Projects without increasing the size of the Evaluation Unit and without converting it into a unit larger than all others operating at the Center. This was achieved by dividing the responsibility for the evaluation activities related to each project into two parts. One part of the evaluation activities was assigned to the Curriculum Project team itself and the other remained the responsibility of the Evaluation Unit. The Evaluation Unit guided the Curriculum Project teams in evaluation and provided some technical services, but the role of carrying out those activities became the responsibility of each Curriculum Project team. Each Curriculum Project team nominated one of its members to serve as the coordinator of evaluation activities within the Project, and this person also served as a liaison between the Curriculum Project team and the Evaluation Unit. The liaison person participated fully in all daily activities of the Curriculum Project team, took part in the team conferences and contributed to the development work of the project, while at the same time carrying out and coordinating evaluative activities. Apart from the current evaluative activities coordinated by the Project Liaison Person, the Evaluation Unit itself sometimes initiated and carried out small-scale evaluation studies related to a particular Project or cutting across several projects facing similar problems. Thus it turned out that in practice both internal and external evaluation activities were carried out with regard to each Project. The internal evaluation was done by the Project team itself, and the external evaluation by the Evaluation Unit. Clearly, the evaluation done by the Evaluation Unit cannot be fully considered as an external evaluation, since it is performed by a team which has continuous and prolonged

contact with the production team. It is nevertheless somewhat detached from the production activity, and is consequently less affected by the bias of the development team than are the evaluation activities initiated by the Liaison Person.

This organizational arrangement permitted very intensive evaluation activities without altering the original balance between the relative magnitude and power of the Evaluation Unit and the individual Curriculum Projects. Given the fact that most Curriculum Projects faced similar evaluation problems, and that the nature of evaluation activities assumed the form of Serial Evaluation Activities, it became relatively easy to provide a series of guidelines common to all projects. A weekly seminar for the Liaison Persons of all the Projects served a double purpose, firstly to guide the Projects in whatever evaluation they were involved in at the time, and secondly to provide systematic knowledge about basic topics in the field of evaluation. Additionally, conferences were scheduled in which the Evaluation Unit team, the team of a particular Curriculum Project and the Liaison Person participated, to discuss unique problems related to the evaluation of that particular Project. If needed, the Liaison Person could recruit help for carrying out evaluative activities from the Curriculum Project team. In many cases auxiliary workers affiliated to the project performed observations in classes or administered tests. Moreover, in certain cases, at some critical stages of the program development, the Liaison Person succeeded in recruiting the whole Curriculum Project team for carrying out evaluative activities for a short period of time. The cooperation between the Evaluation Unit and any single Curriculum Project in carrying out a study is schematically outlined in Table 1. This table presents sequential steps in carrying out a study. The activities appearing in rectangular frames are joint decisions of the two teams. It can be seen that both the determination of the study parameters and the formulation of the recommendations are performed jointly by the two teams, while the implementation of the recommendations is the responsibility of the Project team. The major responsibilities of the Evaluation Unit consist of guiding the work of sampling, instrument development, providing training for people doing the work of data collection, carrying out data analysis, publishing the evaluation results, and filing the data and the instruments used in the studies. But beyond the specific test-oriented activities, the existence of the Evaluation Unit constitutes a permanent reminder of the necessity for empirical examination of the quality of new curricula. Of course a reminder does not in itself initiate action.

It is up to the Evaluation Unit to create cooperation with the Project teams and to motivate them to carry out evaluative activities. Without the Evaluation Unit's vigorous intervention, the effect of a reminder may be no more than that of creating a remorseful feeling by the Curriculum Project team for not doing what one is expected to do.

Curriculum Project Team	Liaison Person	Evaluation Unit
.Identifies problem		.Suggests topic for study
	Agree about parameters of the study	
		.Determines tech- nical features of the study (sampling, ins- truments, etc.)
	.Performs technical preparations	
		.Controls, edits
	.Performs field work	
		.Carries out statistical analysis
	.Interprets results	
	Formulate conclusions, suggestions, recommendations	
.Implements recommendations		.Publishes, files data and reports

Table 1. Sequential Steps in an Evaluation Study.

The description of the formal structure should be complemented by details of some of the informal features of cooperation between the teams. Despite the uniformity of the pattern of the cooperation between the teams, some differences emerged regarding the Liaison Person in the Curriculum Project. Three different patterns emerged.

1. The Liaison Person had had his basic training in evaluation. He had limited knowledge in the subject area of the Project, and was delegated to the Project by the Evaluation Unit.
2. The Liaison Person had had his basic training in the subject area of the Project. He was nominated as a Liaison Person by the Curriculum Project.
3. The head of the Curriculum Project himself served as Liaison Person. This usually happened whenever the head of the Project had some basic training in evaluation and wanted to participate in all decisions related to evaluation. In most cases the willingness of the Project Head to serve as a Liaison Person reflected his sincere interest in evaluation, but sometimes it was a result of his fear or suspicion that evaluation might jeopardize his freedom of action.

While each of the three patterns had certain advantages, the second proved to be the most fruitful pattern of cooperation. It turned out that being immersed in the Project was extremely important for establishing a good working relationship with the Evaluation Team, and facilitated the implementation of the evaluation recommendations.

Contrary to expectations, the the heads of the Curriculum Projects did not function well as Liaison Persons. Their commitment to the Project and the continuous pressure exerted upon them to increase the quantity of instructional materials produced by their teams left them very little time to act in the roles of Liaison Persons. Given their dual alliances, their commitments as heads of the Projects were usually dominant, frequently at the cost of their neglecting their duties toward the Evaluation unit.

Summarizing the experiences derived from the organizational structure of the Curriculum Center one may suggest the following principles as general guidelines for Serial Evaluation Activities operating within the framework of institutes.

1. An autonomous Evaluation Unit, even if it is a small one, may serve as a catalyst for initiating evaluation activities, the limits of which exceed the working capacity of the Unit itself. It may create awareness of the work of evaluation and emphasize its role in the course of the daily activities.
2. Fruitful cooperation and mutual trust between the Evaluation Unit and other units dealing with the target being evaluated can be successfully established when the Evaluation Unit keeps a low profile within the ins-

titute and does not receive privileges, such as access to resources, which are not granted to other units.

3. A realistic assessment of actual needs in terms of manpower and other resources and their satisfactory provision constitute prerequisites for the successful operation of the Evaluation Unit. It is not the absolute size of the budget which determines the successful operation of the Evaluation Unit, but rather the match between the resources available and the appropriate definition of the evaluation tasks.
4. Within each production sub-unit of an institute, the separation of the responsibility for evaluation from the responsibility for running the production unit, may have beneficial effects on conducting evaluation studies and in having the decisions made based on the results. A Liaison Person who is not himself responsible for the production may create better cooperation between the two units than would a relationship which is built solely on personal contacts between the heads of the Units.

A CONCEPTUAL FRAMEWORK FOR EVALUATIVE STUDIES

The organizational structure discussed in the previous section characterizes the Unit which carries out the studies. This section deals with the characteristics of the actual work carried out by the Evaluation Unit. As a working definition of evaluation, the Evaluation Unit adopted the statement that "evaluation is a systematic data collection and data summary process which comes to increase the rationality of decision making". Operating with this working definition of evaluation, three principles are presented in this section. They reflect characteristics of evaluation activities within the Israel Curriculum Center, but in a broader sense they may apply to Serial Evaluation Studies of any types carried out in institutes which are responsible for both the evaluation and the target being evaluated. The three principles are:

1. Since most projects face more decision situations than can be examined by the means of empirical studies during a reasonable amount of time and with limited resources, the Evaluation Unit has to determine priorities and to select topics for study.
2. Since during the process of carrying out a project the team may encounter numerous decision situations, the evaluation activities dealing with a project consist of a series of small-scale focused studies, rather than a single complex one.

3. Though the evaluation of one project may be quite different from the evaluation of another within the framework of Serial Evaluation Activities, it is recommended that a subset of Minimum Evaluation Requirements be defined, which should be incorporated in the evaluation design of each project.

SELECTING TOPICS FOR STUDY

The need of keeping a proper balance between investments in evaluation and in the project being evaluated has already been emphasized in the section dealing with the organizational structure of the evaluation unit. Evaluation has to be carried within a framework which has limited resources in terms of manpower, time, etc. and it cannot consequently afford to address itself to all the questions and issues which may bother the project team in the process of its work. The process of evaluation as specified by Alkin (1970) implies the task of "ascertaining the decision areas of concern", and we would like to add to Alkin's statement the qualification "which most requires support and clarification, based on empirical data collection". Indeed in the process of the work, the project team frequently will have to satisfy itself in handling "decision areas of concern" by consulting experts or even by sharing experiences accumulated during the time. It is impossible to conduct an empirical study whenever one comes across "decision areas of concern". The evaluation team together with the project team have to identify those road junctions which can be crossed without stopping by utilizing intuitive knowledge based on past experience, and those junctions where a full stop is required and where data should be collected systematically before any decisive step is taken.

The selection of topic should be done on the basis of consent between or at least a compromise among the two teams. The suggestion of any topic may nevertheless come from either of the teams. It is true that it is usually the project team which feels confused when arriving at a junction, and that it is accordingly more likely to ask for help in identifying the best road to follow. But it does not mean that this always is so. The project team may be in a hurry, and, reaching a junction, may deliberately close its eyes and act as if only one way would be open before it. Alternatively, it may concentrate on a certain pathway without noticing the existence of alternative ones. In such cases, it is the responsibility of the evaluation team to sketch the road-map exactly and to call attention to the existence of

such alternatives. This should be a cooperative task in which each of the partners has its own well-defined role. The project team has to move forward, and the evaluation team has to put up the road signs at the crucial junctions.

The cooperation between the two teams has to take the form of cooperation between equals, where the nature of the expertise of the two teams lies in different fields. It would be wrong to claim that the evaluation team operates at a higher scientific level than does the project team. The truth is that there are some junctions where the project team's judgement provides valid guidelines for selecting the way to be continued, and other junctions where the evaluator's expertise can provide the most valid guidelines. The evaluator needs to master not only the methodological techniques of producing satisfactory evidence, but also how to acquire insight in locating those junctions where his evidence is most relevant and valid.

A Series of Focused Studies

The development of a new program and the implementation of an existing one are processes which require a relatively long period of time. Institutes tend to operate according to work plans which extend over years rather than months or weeks. The work plan cannot, and should not, nevertheless, constitute a detailed blue-print, which is meant to specify all steps to be taken. Consequently, during the process of carrying out a plan, institutes face the need of continuously making decisions. Accordingly, evaluation units have to conduct studies focusing on a particular decision problem faced by the institute in whose service they operate. One of the most important traits of such studies is the timeliness of the evaluation results. Indeed, one of the major distinctions made between research and evaluation activities is the time dimension of the study.

The traditional conception of scientific inquiry links science to the notion of "timelessness". Science, according to this traditional notion is a search for the eternal truth. Time is not one of its inherent features. On the other hand, evaluative information should be available at the time when the development team faces a decision situation. The pace of the development imposes the criterion of timeliness upon the evaluation and sets the deadlines for it.

To meet the requirement of timeliness within the framework of Serial Evaluation Activities, one should preferably focus on a single decision situation and design short duration studies.

Within the framework of the Curriculum Center it was found convenient to utilize a classification scheme of decision situations based on a mapping sentence of three facets: The stage of the curriculum development, the component of the program or the entity being evaluated and the type of the decision. A brief description of these facets is given below.

The Stages. In the development of a curriculum project six consecutive stages have been distinguished. The focus of development and evaluation activities is described in Table 2.

Stage	Roles of Development	Roles of Evaluation
Determination of General Aims	.Decisions about: . general aims . school structure	.Studies on: . expected changes . cultural values . social forces . present level of achievement . feasibility of programs
Planning	.Writing outlines .Preparing instructional material	.Examining adequacy of objectives, contents, strategies .Judgement of material
Tryout	.Monitoring teaching in tryout classes .Modifying material	.Collecting evidence through observation, judgement, discussion with teachers, students .Student products
Field Trial	.Slightly modifying the program .Determining optimal conditions of program use	.Selecting sample .Collecting evidence about the efficiency of program under various conditions
Implementation	.Links with supervisors, the examination system, teacher training	.Examining final form .Evidence of efficiency of system links .Evidence of efficiency of teacher training
Quality Control	.Implementing recommendations .Planning "second generation" programs	.Examining quality of implementation .Studying reasons for changes in efficiency .Suggesting remedies if needed

Table 2. Six Stages of Curriculum Development and Evaluation (Lewy 1977).

The entity. Essentially, evaluation is concerned with the efficiency of a program as a whole. Thus the basic concern of evaluation is the success of the entire program including all its components. Quite frequently, however, evaluation may deal solely with specific components. The focus of a small-scale evaluation study may be a particular chapter of the program; a particular activity associated with its use, such as the organization of the dissemination network; or a particular type or instructional materials included in it, such as the textbook, the teacher's guide, audiovisual aids, enrichment supplements, experimental equipment, or the teacher-training program. Moreover, an evaluation study may be concerned only with specific features of these components. For example, evaluation may deal with some unique aspect of the textbook such as the quality of the illustrations, the clarity of the explanation, the readability of the text, the sequence of learning experiences, or the adequacy of the exercises. Some unique aspect of the program as a whole may be investigated, such as the efficiency of the type of classroom management prescribed by the program, or the efficiency of a unique learning strategy employed in it.

The Type of Decision. Three major types of decisions are made during the process of developing a program and implementing it: 1) selecting program components, 2) modifying of program elements, and 3) qualifying the use of the program.

1) *Selecting Program Components*

At the early stages of program development, questions arise as to what should be included in the program, what should be taught, what strategies should be employed, and so on. Frequently, several alternatives are presented to the program developer, and he has to select the most appropriate one.

2) *Modifying Program Elements*

At the tryout stage of the program it may turn out that some element (e.g. an exercise, illustration, or explanation) contains certain flaws. Perhaps it does not adequately help the students achieve certain desired goals, or perhaps it is producing undesired effects. Evaluation results may call attention to such problems, and may help program developers improve their program by eliminating the flaws.

3) *Qualifying the Use of the Program*

A third type of decision may be that of qualifying program usage. In other words, evaluation may aid in the specification of the optimal or

minimal conditions for usage. Unlike the other two types of decisions, which deal with issues related to the formation of the program, the third is not related to the amendment of a program, but rather focuses on the task of recommending how and under what conditions a program should be used.

Conditions may be specified for various aspects of a program. Specifications may be given about teacher training; availability of equipment, space, time; professional prerequisites; the consent and support of the community or the broader environment; existence of a skilled supervisory or monitoring staff; and so on. For any particular school, the specification of conditions may constitute a basis for the selection or rejection of a particular program.

A Classification Scheme. The three facets described above combined in a mapping sentence may serve as a classification scheme of decision situations which should be dealt with in evaluation studies.

The mapping sentence in its totality constitutes an overall inventory of decision situations which one may encounter during the process of curriculum development. Each such decision situation may require the conducting of a short-duration focused evaluation study, in which the summary of empirical data may contribute to the rationality of decision-making.

The mapping sentence containing the three facets is presented in Table 3.

Evaluation is the provision of information at the concerning the sake of making decisions about	A: STAGES	
	determination of aims planning tryout field trial implementation quality control	stage of program development
	B: ENTITY	
	teacher's guide study material equipment whole package	for the
	C: DECISION SITUATIONS	
	selecting elements of modifying qualifying the use of	the program

Table 3. Mapping Sentence Classification Scheme of Curriculum Evaluation Studies.

One may define a particular type of focused evaluation study by selecting a single line from each of the three facets appearing in the mapping sentence. Thus, for example, a focused evaluation study may aim to provide information at the tryout stage of the program development concerning the teacher's guide for the sake of making decisions about modifying the original version.

Minimum Evaluation Requirements

It has been noted that the evaluation of a particular educational program is not performed through a single study. At various stages of program development alternative pathways of action may appear, or doubts may arise concerning the merits of some parts of the program. In such cases the evaluator may be asked to define a problem area precisely, to collect appropriate data and to summarize them in a way that may help decision-making. Thus, each sub-study concerned with the evaluation of a program focuses on a unique, well-defined problem area. Two such sub-studies may differ from one another from the viewpoint of the focus of the investigation, type of data collected, and so on.

It has been emphasized that it is the evaluator's task to identify the crucial decision situations and the critical junctions where evaluation may yield the most useful information. These junctions may differ from one program to another. Nevertheless, whenever an evaluation unit assumes responsibility for carrying out Serial Evaluation Activities, one may determine within the evaluation space several junctions where the evaluator is requested to place road-signs on the basis of his summary of systematically collected empirical data.

Within the framework of the Israel Curriculum Center, three such junctions have been identified, and at each of these junctions empirical evidence of a specific type is requested in order to determine which path to follow. These junctions are:

- 1) Obtaining expert judgement about the scientific quality of the instructional materials as soon as the first version of the program has been prepared.
- 2) Observing the teaching and learning process during the first tryout of the materials in the class in order to determine the feasibility of using these materials and to specify teacher-training needs.
- 3) To examine the cognitive achievements of the learners at the end of the first tryout of the materials in the experimental classes and to compare

the actual level of the achievement with the expected level of achievement as specified by the program objectives or implied by the contents taught.

These three evaluation activities have been considered as minimum requirements by the Evaluation Unit, with the first two activities performed mainly by the Project team under the leadership of the Liaison Person and under the supervision of the Evaluation Unit, while the third one was mainly performed by the evaluation team.

THE MANAGEMENT OF EVALUATION STUDIES

It has been already mentioned that evaluative information can serve as input for decision-making only if it needs the requirement of timeliness. Stufflebeam (1969) listed the trait of "timeliness" as one of the most important of the traits which determine the merits of an evaluation study, and he attributes to the "timeliness" a significance matching that of reliability and validity. Timeliness in evaluation studies is to a large extent a function of realistic planning and good management. Fortunately, evaluation units operating within the framework of institutes and carrying out Serial Evaluation Activities can more easily meet the requirement of timeliness than can *ad-hoc* free-lance contractors. The former may take advantage of the fact that they are dealing with events or targets which repeat themselves according to a recurrent pattern.

The Evaluation Unit of the Israel Curriculum Center established an infrastructure geared especially toward the needs of the Serial Evaluation Studies carried out in the Center, and which facilitated the quick completion of focused studies. In this respect the Israel Curriculum Center tried to capitalize on experiences accumulated in various institutes of survey research. Indeed in various fields of human behavior very quick and efficient data gathering and data analysis techniques have been developed. Institutes have specialized in producing well-organized information for their clientele about important issues. They have extended their services to areas such as predicting voting trends, examining public opinion about dominant political and social issues, carrying out market research, exploring exposure to mass communication, efficiency of advertisements, etc. The Israel Curriculum Center applied some techniques developed in the domain of surveys for enhancing timeliness. One has to admit that there are differences between the task of carrying out surveys of the above-mentioned types and evaluation studies in

the field of education. Most surveys concentrate on well-defined behavior patterns and content themselves with reporting the distribution of observable phenomena. They rarely face the problem of a search for criteria or of implementing conceptually defined variables. They are not usually asked to forward explanations of observed behavioral patterns and changes. There are nevertheless procedures employed in surveys which are transferable to educational evaluation. Additionally, there are developments in the field of evaluation which, when combined with the routines of the survey techniques, can enhance timeliness.

We will discuss below some features of conducting evaluation studies which may enhance timeliness.

Comprehensive Conceptual Framework

Conducting short duration, focused studies does not imply that one may disregard the need for establishing an adequate conceptual framework as a basis of formulating questions, selecting variables, determining the mode of data analysis, etc. By no means should evaluation be degraded into a series of data collection and data analysis routines. Without a conceptual framework, data summaries remain disjointed sets of facts, which do not contribute to our understanding of the dynamics of various phenomena. But in a serial evaluation study the establishment of the conceptual framework should be done in a comprehensive way and commonly for the entire series of studies, rather than separately for each one. One may identify several basic concerns, concepts or variables, etc. which play a crucial role in the whole series of the studies and then in each particular study focus on those which are uniquely relevant. Thus, two studies of a single series, though founded on a common comprehensive conceptual framework, may differ from each other with regard to their actual design.

Permanent Institutional Setting

Surveys are conducted by professional people who are permanently employed in institutes carrying out such work. Although surveys are less complicated professionally than are evaluation or research studies, survey institutes are often staffed by more highly qualified and competent professional people than are evaluation studies. No doubt the permanent nature of their job is a crucial factor in attracting highly qualified people to those institutes. Experts are not hired to perform a particular survey, but are rather asked to perform a series of studies, which have several recurrent components in

common. Thus, the experts may develop a high level of familiarity with the working procedures relevant to their job.

While similar arrangements do exist in the field of evaluation and several institutes also employ permanent staff members, it is less typical for evaluation studies to be carried out by professional people whose full-time work is devoted to a single study.

Timeliness in evaluation can be enhanced by increasing the number of experts having full-time jobs on a single evaluation project. In this respect, the working pattern of the television news production team may serve as a model for evaluators. The television team knows that its work has to be ready at a certain time, otherwise it will be completely disregarded. The demand to meet deadlines, at least in television news production, does not necessarily diminish the quality of the programs.

Pre-Arranged Sample Plan

One of the time-consuming procedures in the process of carrying out an evaluation study is drawing a sample and ensuring the cooperation of the sampled subjects. In studies which deal with whole populations, the question of sampling does not exist. But the majority of evaluation studies in education will restrict data collection to a sample of the target population. Drawing a sample requires a series of sequential steps, each of them quite time consuming: the exact definition of the target population, the identification of those stratifying variables which are the most effective in reducing the magnitude of sampling errors, etc. Sampling in educational studies is especially complicated because it usually requires multi-stage sampling procedures, e.g. the selection of schools, classes within the schools and individuals within the classes. Furthermore, some variables describe individual behavior, such as the learner's scores on certain cognitive or affective measures, while other variables relate to larger units, such as the class or the whole school. Examples of such variables are the training of the teacher, the utilization of a certain curriculum in the school, the availability of equipment of a certain type, etc.

In Serial Evaluation Activities the sample plan should be prepared in advance, before the actual start of a particular inquiry. This demands a sequence of activities as described below:

List of the Population: The existence of a full list of subjects included in the target population enhances the process of drawing a sample considerably. In educational evaluation, no matter whether the subjects of the study are pupils, teachers, classes, or schools, the availability of a list of schools constitutes a sufficient starting point for drawing samples. In many countries, where compulsory education has been introduced, such lists can be obtained from some central agency. Where no single central agency can produce a full list of all schools, usually a research institute can obtain partial lists from regional or sectional school authorities. In educational systems where such full lists cannot be readily obtained as a first step for running Serial Evaluation Activities, one should compile a list of schools which represents fairly all the types of schools in the system. This list should be larger than the actual sample needed for a particular study. Ideally, it should be at least five to six times as large as the sample which one intends to approach in a single study.

Indices of Stratification: While adequate indices of stratification may vary from system to system, or even within a single system from problem to problem, the experience accumulated in national and international evaluation studies has shown that the list of such indices is relatively limited. They include such variables as: geographic region, ethnic and linguistic composition of the learners, socio-economic position of the learner's parents (often measured by the parents' occupation or education), sex of learners in school (boys, girls, or coeducative schools, the proportion of boys and girls), type of school (comprehensive, selective, characteristics of those pupils selected for the school), ownership (state, public, private, etc.). Additionally, from the point of view of drawing a sample, the most important issue is probably the size of the school. The size of the school will determine whether further sampling within the school will be needed or, alternately, whether all subjects (learners, teachers, etc.) within the school will be included in the sample.

Regional Liaison Persons: Carrying out a good sampling plan requires more than the selection of an appropriate sample. There is a need to assure the cooperation of the subjects in the sample. In hierarchical organizational structures, it may be necessary to obtain the permission of people at various levels of the hierarchy, in addition to the consent of the subjects themselves. In some systems the superordinate authorities may issue an order to the lower order bodies to carry out the data collection request. But

data collection requires more than giving permission, issuing orders, or expressing consent. There is a need to create a climate of mutual confidence, an understanding of the aim of the study, and an appreciation of it. To attain these, there is a need for continuous contact with the schools included in the sample. It may well be that the contact itself constitutes a certain contamination effect. Nevertheless, without contact there is a danger of careless data collection or of subjects dropping out of the sample. The selection of the sample and the contact with the representatives of the schools should be established before the specific goals of a particular study have been fully operationalized, let alone before the instruments to be used are fully developed.

Utilization of Item Banks

Another technique which may enhance timeliness is the utilization of item banks. Item banks contain test or questionnaire items which were developed by experts, and which were empirically tested. The items contained in the bank are characterized by a series of parameters. The most commonly used parameters are content classification, the mental processes involved in answering; the difficulty level, using a variety of indices; the popularity of different distractors; difficulty indices for various sub-populations; the relatedness of the item to other items in the bank, or to a series of items. Item banks have been established in several countries in the last decade, among them Great Britain, the Netherlands, New Zealand, Germany, and the U.S.A. Stanford University created an item bank of statements describing teachers' behavior. The International Association for the Evaluation of Educational Achievement (IEA) recently developed a plan for an international item bank (Postlethwaite, et. al., 1979). The utilization of item banks has nevertheless not yet become wide spread.

In the systems where its utilization reached a peak, it served only a negligible percentage of the total testing activities going on in the system.

There are several ways of managing item banks. In the simplest setting, the compilation of the bank is done by an expert team. The items themselves, with their descriptive parameters, are then published and disseminated for open use. The item bank in arithmetic developed in the Netherlands, constitutes an example of such management. Similar arrangements were made in New Zealand. A higher level of central involvement in maintaining item banks is an arrangement by which the items themselves are not published for the general public, but are kept in the files of the central agency. The central

agency distributes descriptions of the item bank, and the clients order what they need by specifying the basic parameters of the items to be included in the text. The central agency provides the tests, collects the answers of the respondents and performs the analysis.

In practice, one may observe many combinations of the techniques described above. Thus, in many cases, the item bank is published and distributed, each client makes up its own test and administers it, and the central agency performs the scoring and data analysis.

Tests can be obtained from the item banks directly by utilizing computers. Computer-aided test construction may reduce inter-student cheating since it is capable of producing many varieties of item and response sequences for a single test. Thus, for example, for a 30-item test, each item having five alternative responses, millions of different forms can be arranged. This means that 30-40 different forms of the test may be sent to a single school or institute. The utilization of large-scale computer-managed item banks would require the following actions:

- 1) Absorption of new items. An item bank should never be considered closed. Otherwise, even a large-scale item bank becomes exhausted and creates undue interest in particular items.
- 2) Development of routines for random selection of items. An item which has been used in a particular test should have a lesser but still non-zero chance of appearing in another test.
- 3) Development of routines for printing the same tests in different item and response sequences.
- 4) Permanent updating of item parameters on the basis of results obtained by recurrent administration of the items.

Two developments in the domain of testing may increase the effectiveness of using item banks: matrix sampling and sample-free calibration of item difficulty indices.

Matrix Sampling: The technique of matrix sampling developed routines by which one may randomly assign different samples of items to different individuals participating in an evaluation study. Let us take the example of an evaluation study in fourth grade arithmetic, in which a sample of 800 students was selected. Given a one-hour testing time for each student, one should prepare a 40-item test. Now, utilizing the technique of matrix sampling, one may

create a variety of tests, each one utilizing a different set of 40 items, and thus collect responses from the learners not on 40 items, but on a total of 200-250 items. Despite the fact that each item might have been answered by only 200-300 students, reliable and valid estimates can be attained about the performance level of the whole population on 200-250 different items.

Sample-Free Calibration: Another recent development in the field of testing is a technique for determining the difficulty index of an item without regard to the general ability level of the sample of the students which answered the items. In classical test analysis procedures, the difficulty index of an item was determined by administering it to a sample of learners, and the percentage of the correct responses served as the parameter of the difficulty level. But several critics of this approach claim that by such derivation the difficulty index becomes a function of the ability level of the students. If an item was included in a test given to a group of low-ability students, its difficulty index would be higher than if the same item was given to a group of higher-ability students. To handle these problems, sample-free test calibration methods were developed (Rasch, 1960; Choppin, 1978; Bauer et al., 1977; Kempf, 1975). Using this method, one may derive difficulty parameters which are not related to the ability level of the respondents in the sample. One great advantage of this method is that it enables one to derive comparable scores for tests composed of different sets of items. Combining the sample-free item calibration processes with the matrix sampling technique is one way to enhance the timeliness of providing comprehensive information about topics of interest.

Advance Planning of Computer Analysis

With the advent of standardized computer programs, one would expect that the time needed for performing analysis of data would be largely reduced. No information is available on whether or not time reduction has indeed occurred but first impressions of a variety of data summary projects seems to indicate that the time necessary for data analysis was not shortened. The availability of easily manageable computer programs decreased the amount of thoughtful planning of data analysis and increased the appetite for a greater range of data analysis modes. In most studies, only a small portion of obtained data summaries are reported. Frequently, such a selection of "results" not only causes prolongation of the time needed to complete a project, but actually causes a bias, resulting from the selection procedures.

In Serial Evaluation Activities, careful planning should prescribe the major data analysis procedures and the forms and the patterns of presenting results. By doing this, one may develop computer outputs in a photo-proof form for immediate inclusion in the research report. An example of such an output set is a test-scoring program produced for the University of Illinois Medical Center (Lewy and Crawford, 1966), which contained tables and histograms in a form which was considered useful for the evaluation reports. To ensure timely data summary, it proved advantageous to perform a "dry run" analysis, in which a small set of simulated data were analyzed, in the exact form in which they were to appear in the final report. In general, the inspection of the "dry run" results permits the early detection of errors in the control card prepared for the summary of a particular set of data.

THE CLIENTELE

Evaluation experts suggested that each evaluation activity should be geared towards the needs of a specific clientele (Stake 1972; Stake and Craig 1974). Evaluation methods and reports should be not only relevant but should also appear credible to the decision-maker. The results should be presented in a way which yields easily and unambiguously derivable decision rules. Moreover it has been recommended that the decision rules should be determined on the basis of an agreement between the evaluator and his client before the study design has been finalized. Some experts claim that if agreement about the decision rules cannot be obtained before starting the study, there is little chance that the evaluation results will affect decision-making. One should nevertheless be aware that while agreement in itself may be a necessary condition, it is certainly not a sufficient one. In science, no agreement has validity, unless it is based on publicly defensible principles. In any scientific endeavors the parties are not only permitted, but are supposed to cancel agreements, if one of the parties realizes that the agreement was based on erroneous or non-valid principles. The notion of "publicly defensible" implies that it is not enough for an evaluation study to be an internal affair among contracting parties, but it should be made accessible to a non-homogeneous type of clientele. This leads to the formulation of two principles:

- 1) Evaluation results should be presented to a multiple clientele.
- 2) Evaluation results should be presented in a form which enables access of the clientele not only to the recommendations but also to the data on which the recommendations are based.

Multiple Clientele

The organizational setting of the Evaluation Unit and the definition of evaluation as a series of short-duration focused studies clearly suggest that the direct clientele of the Serial Evaluation Activities carried out at the Israel Curriculum Center is the Curriculum Project team, which is responsible for producing new curriculum materials. It is supposed to utilize the evaluation results for the purpose of decision-making. The reports produced by the Evaluation Unit should nevertheless be valid for clienteles of different types too. Firstly, they should contain valuable information for the superiors in the office hierarchy of both the project team and the evaluation team. This constitutes a safeguard that the project team is not in a position to disregard evaluation results. It is not that one fears a sinister attitude of the project team toward evaluation results. But in an institutional setting one cannot build a system in which the execution of a certain task is left entirely to the discretion and benevolence of those who have to carry out that particular extra work. There is need for a certain control mechanism. This may be achieved by broadening the clientele of the evaluation report. It may well be that the form of the report presented to one's superiors should be different from that presented to the project team, but it is important that the two parties should be informed about the results. Of course this can be done only if mutual trust exists among all parties interested in the project, so that the report does not constitute a threat for those whose work is evaluated. This is especially important in a situation where an evaluation team maintains continuous working relationship with the project team during a period of several years.

Additionally, evaluation results should be communicated to the community of evaluation experts. Their approval and their acceptance of the decision recommendations serves as support for the demand that the recommendations should be implemented. Finally, teachers, parents, etc. should also be informed about the merits and shortcomings of programs suggested to them.

Access to Data and to Written Reports

Evaluation experts tend to denigrate the importance of preparing formal evaluation reports, let alone reports which are written in a scholarly style and which contain detailed analysis of all the data collected within the framework of a study. They often consider it quite satisfactory if in certain cases evaluation results are transmitted orally. Indeed, the decision-maker can often better absorb an oral communication than a written report.

The preparation of a written report is a time-consuming process and from the point of view of the decision-maker may be a waste of time and money. In practice it may be almost impossible both to prepare scholarly evaluation reports and to complete evaluation studies within a reasonable time.

While one may agree that in the process of curriculum evaluation constant contact should be maintained between evaluators and decision-makers, and that research findings should be conveyed as soon as possible, there are several reasons which support the contention that formal evaluation reports should be prepared as frequently as possible.

First of all, an orally or informally communicated finding is less likely to affect action than a formally written statement. The decision-maker may more easily disregard or "explain" an informally conveyed message than a research report. This is likely to happen when following evaluation recommendations requires extensive work on the part of the curriculum worker. Secondly, the standard of curriculum evaluation may deteriorate if no demand is imposed upon the evaluator to prepare scholarly and defensible reports. Finally, one has to consider the fact that curriculum evaluation is an emerging field of scientific endeavor. A sound theory of curriculum evaluation can be developed only on the basis of the critical analysis of well-written published empirical studies. The lack of empirical studies in this field may create a situation in which theory and practice are not related to one another at all.

REFERENCES

- ALKIN, M.C. Products for Improving Educational Evaluation. *Evaluation Comments*, 2(3), 1970, 1-15.
- BAUER, M., KEMPF, W.F., POSTLETHWAITE, N. and LEWY, A. Collecting and Analyzing Evaluation Data. In: M.A. Lewy (Ed.) *Handbook of Curriculum Evaluation*. New York: Longmans, 1977.
- CHOPPIN, B. Item Banking and the Monitoring of Achievement. Slough. (England): National Foundation for Educational Research in England and Wales, 1978. (mimeograph).
- KEMPF, W.F. A Dynamic Text Model and its Use in Microevaluation of Instructional Materials. In: M.H. Spada and W.F. Kempf (Eds.) *Structural Models of Thinking and Learning*, Beru, Haber, 1976.
- LEWY, A. (Ed.) *Handbook of Curriculum Evaluation*. New York: Longmans, 1977.
- LEWY, A. and CRAWFORD, R. Scoring Text Battery. *Educational and Psychological Measurement*, 26, 1966, 185-88.
- POSTLETHWAITE, N. The IEA Item Banking Project Proposal. Hamburg: University of Hamburg, Dept. of Education, IEA, 1978 (mimeograph).

REFERENCES (Cont.)

- RASCH, G. *Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research, 1960.
- SHAKE, R. *Responsive Evaluation*. Urbana: University of Illinois, 1972 (mimeograph).
- SHAKE, R. and CRAIG, G. An Evaluation of TCITY, The Turin City Institute for Talented Youth 1971. In: Stake R. (Ed.) *Four Evaluation Examples*. Chicago: Rand MacNally, 1974.
- STUFFLEBEAM, D. Evaluation as Enlightenment for Decision Making. In: M.W.H. Beatty (Ed.) *Improving Educational Assessment*. Washington, D.C.: Association for Supervision and Curriculum Development, 1969.

CONTINUOUS EVALUATION AS AN INTEGRAL PART OF CURRICULUM DEVELOPMENT

The Evaluation of the High School Biology Project at the

Israel Science Teaching Center

Pinchas Tamir

INTRODUCTION

In his paper on "Continuous Evaluation Services attached to Large-Scale Action Projects" Arie Lewy (1980) describes the activities of the Evaluation Unit of the Israel Curriculum Center (ICC) established by the Ministry of Education and Culture in 1966. As far as science is concerned, the ICC assumed responsibility for producing new curricula and innovative instructional materials mainly for the junior high level (grades 7, 8, 9) while the development of high school and elementary school programs and materials has remained the responsibility of the Israel Science Teaching Center (ISTC) which began its activities in 1963 as separate disciplinary (mathematics, physics, chemistry and biology) projects located in institutes of higher education. This division of labor between the ICC and the ISTC has been maintained to 1980 and is likely to continue at least in the near future. While science project teams at the ICC and the ISTC maintain close relationships, different modes of operation have emerged, not only between the ICC and the ISTC but also between different projects of the ISTC. In this paper the writer intends to focus on evaluation as operated in relation to the Israel High School Biology Project (IHBP). This particular project was selected because of the special attention that has been paid by its staff to evaluation as well as the major role that evaluation has played in its day-to-day operation. As a member of the development team from 1964 to 1966 and as the director of the IHBP since 1968, the writer is also more familiar with events related to this project.

In his account of the operations of the Evaluation Unit in the ICC, Arie Lewy emphasizes the following:

1. Since the ICC was set up with the aim of introducing curriculum reform in the schools, its major activity has been the production of innovative programs and instructional materials; hence, evaluation was designed to aid and serve this process of production and to help in improving the quality of the products.
2. The establishment of a separate Evaluation Unit helped to maintain its autonomy and independence. Yet, its location within the ICC enhanced close cooperation with the different curriculum development units, and helped the evaluators in locating critical junctions where evaluation evidence would be most useful.
3. Evaluation studies were initiated either by the members of the curriculum development projects or by the Evaluation Unit. Contacts between the two were maintained by project members who acted as liaison persons. The actual selection of topics for study was made by the mutual consent of both groups.
4. Since most curriculum projects faced similar evaluation problems, a standard procedure of Serial Evaluation Activities aimed at six stages (Lewy, 1977) was developed and implemented. It resulted in small, limited, focused studies which provided specific information to help in making decisions at critical stages. A set of minimum evaluation requirements was also established. These requirements included expert judgement of the quality of the instructional materials, observing tryout classes and assessing students' cognitive achievement at the end of the first trial against stated program objectives.
5. The major responsibility of the Evaluation Unit was in guiding the procedures of sampling, in instrument development and in training people in the collection, analysis and reporting of data, which facilitated the quick completion of focused studies.
6. The existence of the Evaluation Unit constituted a permanent reminder to project teams of the need for conducting evaluations and for examining the quality of the new curricula.
7. Fruitful cooperation and mutual trust between the Evaluation Unit and development projects has been successfully established, presumably due to the low profile kept by the members of the Evaluation Unit.

8. The Serial Evaluation Activities of the ICC were geared mainly to the curriculum development teams who were to use the evaluation results for decision-making, and to a lesser extent to their superiors. Timeliness required that many reports be made informally, either orally or in draft form, yet the need for systematic formal reporting to a multiple clientele has been recognized.

It is interesting to compare the evaluative procedures of the ICC as just described with the evaluation strategy described by Leinhardt (1977), as employed by the Learning Research and Development Center (LRDC) at the University of Pittsburgh.

The essence of the LRDC strategy is to incorporate the work of individuals from diverse backgrounds who are working in problems in their own discipline areas (i.e. psychology, sociology, philosophy, anthropology) into an overall program of evaluative research. This strategy is offered as an alternative to the traditional in-house one-shot studies conducted by a single team of evaluators. The emphasis in this strategy is for researchers to satisfy themselves with regard to the scientific excellence of their work and to seek answers to questions which they generate, rather than to attempt to second-guess the consumers' questions. Admittedly, there is no assurance that the answers to a collection of questions generated by different researchers will provide adequate information about the value of the program. Nevertheless, Leinhardt presents convincing examples of studies conducted under this framework which provided useful information to consumers about the nature of the program, the soundness of the formative evaluation (which was conducted mainly by the curriculum developers) and the effects of the program on the learners and on the learning environment. Under this framework the role of the evaluator is to recruit researchers, to coordinate their work and to see that the results are presented in a usable form.

Unlike the ICC and the LRDC, most curriculum development projects in the U.S.A. and in the U.K. were financed only for limited periods ranging between three and five years. In spite of their limited life span, some projects such as Science 5-13 in the U.K. or Harvard Project Physics in the U.S.A. and AESP in Australia incorporated evaluation as part of their on-going development processes. Within these projects evaluation was meant to serve two major functions:

- a) the collection of data to assist in decisions about development and adoption;
- b) construction of instruments to assist teachers in the ongoing evaluation of students in their classrooms.

The Harvard Project Physics presents a unique example of success in employing true experimental design in assessing the effects of its program when compared with that of other programs (e.g. Welch & Walberg, 1972).

Generally, evaluation projects located at universities, whether short-time or continuous, have tended to integrate research with evaluation. Rather than focusing on a series of small-scale evaluation activities following standard procedures as characterized by the ICC model, they have tended to employ innovative procedures, to deal with a variety of outcome variables in addition to cognitive achievement and to develop innovative evaluation instruments.

Typically, the results of research-oriented evaluation studies were published in the professional literature and certainly added interesting and important dimensions to our knowledge about curriculum development and implementation. Unfortunately, the direct impact of these evaluations on decision-making have not been documented. The writer suspects that in many cases there has been no such direct impact simply because most information became available too late.

The Israel High School Biology Project (IHBP) is an example of a project located in a university and directed by university faculty. At the same time, however, it has maintained direct contacts with schools and the actual monitoring of instruction by establishing a variety of services such as supply centers, inservice training, and guidance in schools. Most importantly, it has maintained control over evaluation by undertaking the task of producing matriculation examinations. A brief account of the history of the IHBP will help in understanding the role that evaluation has played in its activities.

Curriculum Development and Implementation

In 1964 a group of selected Israeli high school biology teachers, under the direction of Professor A. Poljakoff-Mayber (a prominent plant physiologist deeply interested in education) undertook the task of producing a new high school curriculum in biology. The first matter to be resolved was whether to produce an original curriculum or to adapt one from abroad.

The decision to adapt was based on four considerations. First, the need for curriculum reform was urgent, while expertise in curriculum development was scarce. Producing an original curriculum, although desirable, would take too long; adaptation promised a shortcut. Second, Israel's financial resources are limited. The Israeli project could not have hoped to produce

an original curriculum comparable in quality to those designed by the BSCS in the United States and the Nuffield project in England (1965a, 1965b), where huge expenditures had been required to secure the necessary time and expertise. Adaptation promised a better curriculum in the end. Third, the principles of biological inquiry are the same in Israel as in the United States and Britain. In 1964, the two main lines of revision in science curricula were teaching the structure of the discipline and promoting inquiry-oriented laboratory work. There seemed little to be gained from a reworking of the disciplinary content. Fourth, the approach of the BSCS and of Nuffield appeared sound, as did the materials they had developed. Adaptation of these programs seemed worth trying. The BSCS was chosen over Nuffield because the BSCS materials were then at a much more advanced stage of development.

The BSCS program is available in three versions (known as Blue, Green, and Yellow), each proceeding from a different point of view. The Blue Version (BSCS 1963a), which employs biochemical concepts as a starting point, was judged too sophisticated for most high school students in Israel. Because of its ecological approach, the Green Version (BSCS 1963b) employs a great many examples from North American environments; to adapt it would have required substituting plants and animals from Israeli environments. At the same time, the Yellow Version (BSCS 1963c) offered two distinct advantages: 1) It was most like the conventional curriculum then in use, and so required the least extensive changes in teaching practices; it promised the least painful change-over, and the most likely success. 2) The expertise of the project head was most relevant to the Yellow Version. Thus the Yellow Version was adopted.

The curriculum development group divided itself into pairs, each of which was assigned a chapter of the Yellow Version. A critical analysis was conducted of the content, examples, diagrams, laboratory exercises, and other features of the chapter. The analysis focused on two questions: To what extent were the various features of the chapter scientifically sound and up-to-date? To what extent were they suited to local conditions? The operation was not just an adoption but an adaptation.

Once every three weeks the whole group met in a central place. In each of these meetings one or two pairs of teachers presented their adapted chapters and a general deliberation took place. In most instances, biologists working on the topics under discussion were invited to participate, and they were the only professionals other than the teachers to enter these discussions. Consequently the decisions made were based chiefly on the opinions of the

scientists, whose competence was highly regarded and whose judgement was therefore usually accepted; the considerations of subject matter dominated the decisions. After these meetings the material was revised and handed to a professional writer to ensure stylistic coherence. The first two parts of the Yellow Version, "Unity" and "Diversity", were prepared in this manner; the third part, "Continuity", was prepared in a later year.

The group decided to publish the text in eight softcover books rather than in one thick hardcover textbook. This decision was a result of two independent considerations:

1. It had become clear that the preparation would take more than a single year, yet the first classes were to start studying the course in 1965. It was thus necessary to adopt a piecemeal publication strategy so that some of the earlier chapters could already be in use, even though later chapters were still in preparation.
2. Students in Israel carry their books home and bring them back to school daily. Heavy textbooks are thus very inconvenient, especially since biology is taught only three periods a week throughout the ninth and tenth grades, and the course was therefore planned to extend through two years. If a single text had been produced, students would have had to carry it for two years, even though only half of it would be used each year. That was obviously undesirable. As it happened, "Unity" was published in 1965, "Diversity" in 1966, "Heredity" in 1967 and "Evolution, Ecology and Behavior" in 1969.

What about evaluation as related to curriculum development? As in the ICC, due respect was given to evaluation within the framework of the IHBP, and provision was made to guarantee its autonomy within the project. This was achieved by nominating an evaluator who had no responsibility in curriculum development but who occupied himself solely with designing and carrying out evaluation. In fact all the decisions described above about the selection of the BSCS Yellow Version for adaptation and the actual procedures of development which followed including the determination of general aims, the planning and actual preparation of instructional materials (i.e. the two first stages of curriculum development described by Lewy, 1977) were not based on either the formal or informal evaluation of the nominated evaluator. These decisions were based primarily on the assumption that a program which had been found to be feasible in the U.S.A. had high probability of being feasible in Israeli schools. Other decisions concerning the selection

of program components, modification of program elements and the specification of minimal conditions for usage (such as minimal laboratory facilities) were made by the development team based on deliberations and expert judgment.

At the same time the evaluator occupied himself with designing a formal evaluation scheme which was aimed mainly at assessing students' achievement in the following areas:

- cognitive achievement in biology;
- understanding the nature of science and its processes;
- ability to analyze and comprehend biological research;
- inquiry skills (such as formulating hypotheses and designing controlled experiments);
- attitudes toward science and nature.

A considerable amount of effort went into the selection and design of evaluation instruments. Some available instruments were translated into Hebrew with some modification (e.g. the Test on Understanding Science - TOUS), some made use of test items taken from tests designed by the BSCS in the U.S.A. and some were designed by the evaluator with the consent of the project team.

The American BSCS Project conducted large-scale evaluation studies until 1963. The conclusions of these studies were based mainly on teachers' reports, staff visits to classrooms, and the results of paper-and-pencil multiple-choice tests given to students. A limited budget had been allotted to evaluation, and it proved rather difficult to obtain the cooperation of professional evaluators (Grobman 1969). Consequently, there has been no systematic evaluation of the BSCS program since 1963 (Grobman 1970).

In the Israeli project, by contrast, evaluation was established as a long-term component. As early as 1965 it had been decided to conduct a follow-up study that would take four years and include three groups of students in grades nine and ten, representing the three types of schools in the country. Based on a battery of tests of cognitive and affective outcomes, this study was finished in 1969. The follow-up of those students who had participated in this study, and who had then continued their study of biology in grades eleven and twelve (as biology majors), was continued through 1971. The results of these studies have been published (see Jungwirth 1969, 1970, 1971, 1972; Tamir 1972 and Tamir and Jungwirth 1975).

Along with the formal evaluation study just described, informal evaluation took place based on visits to schools, discussions with teachers and other sources. A special characteristic of the IHBP was its step by step publication of materials, which extended over five years. This continuous process had its drawbacks. For example, in several instances a class completed the published materials before the next unit was available and the teacher had to continue even though the students had no textbook.

In retrospect, though, the step by step Israeli process has had certain advantages, prime among them the opportunity to learn from mistakes. Feedback collected on the first chapter assisted the team in making decisions with regard to the writing of the next ones. The extended period of development also gave teachers and schools the opportunity to try some of the materials before committing themselves and their students to the new program. This process of consumer evaluation has received little attention by the evaluation experts, yet it may be a decisive factor as far as program adoption is concerned.

By consumer evaluation we refer here to a process whereby teachers and schools utilize curriculum materials as a means for deciding whether or not to use a program, and in what ways to use it with different students and by different teachers. This is an informal evaluation process which takes advantage of the wisdom and judgement of practicing teachers rather than that of learned evaluators.

Diffusion and Implementation

One of the criteria used to judge the success of a program is the rate of its adoption by the schools (Welch 1968). In spite of the existence of a centralized educational system in Israel, the project team decided, in cooperation with the Ministry of Education, that adoption of the program would be purely voluntary on the part of teachers and schools. In the first years, book and materials were made available only to teachers who had participated in a special in-service training course and only to schools which had an adequately equipped laboratory. (The conventional curriculum in biology remained acceptable and operational as late as 1975). These measures were helpful to the teachers and to the individual schools, and at the same time increased the probability of successful implementation of the program.

Naturally, those teachers who participated in the writing group taught the first twelve classes, comprised of 300 students, in the year 1965.

Of these, four classes were in kibbutz schools, one class in an agricultural secondary school, and the rest in city academic schools in various parts of the country. The number of classes and students increased rather slowly in 1966 (550 students) and 1967 (1,000 students). Many teachers were suspicious and chose to wait and see what would happen. Apparently, by 1968, the feedback coming from the trial schools was favorable enough to start a wave of adoption (4,000 students in 1968, 8,000 in 1969, 40,000 in 1974). It is interesting to note the characteristics of the teachers and schools that first chose to follow the new program. In 1969, of about one hundred teachers, seventy had teaching experience of less than three years. In that year about half of the kibbutz schools, 20 percent of the academic city schools, and 5 percent of the agricultural schools used the program. Most of the prestigious city academic schools had not yet adopted it. These were anxious about the new matriculation examination that was to be introduced, and were still waiting for the results of the evaluation study which had begun in 1966. Then, too, some teachers were afraid of the extra work and of the new competencies required to teach the new program adequately.

Later on, after the first two groups of biology majors had taken the matriculation examination and had proved that they could meet its rather ambitious requirements with no damage to their grades, more and more schools joined the new program, with 80 percent of the schools using it in the year 1975, and most schools in the year 1980.

Unlike curriculum projects which have conceived their primary function to be the design and development of curriculum materials, the Israeli Biology Project has placed at least equal emphasis on providing the necessary conditions for adequate implementation and on assessment and evaluation. In each of these phases the Project team has made a point of cooperating with various departments and key persons in the Ministry of Education, as well as with other persons and organizations related to biological education. Such cooperation has proved to be instrumental in the successful implementation of the project. In a country with a centralized educational system, such as Israel, a policy of involving the system will yield better results than one which attempts to fight it.

How does one go about involving the system? The approach used was to identify key persons, such as the Chief Inspector and the Head of the Examination Department, and to secure their interest and sympathy on the one hand, and to seek their advice and opinions on a regular basis on the other. More-

over, where and when necessary, the Israel Science Teaching Center has been ready to assist these government officials in various matters and to respond positively to their ideas and suggestions.

The Utilization of Evaluation Findings

There were eight major findings of the evaluation. 1) Students studying the new program achieved more than did their controls (who studied the conventional program) on paper-and-pencil tests that assessed functional knowledge in biology, comprehension, application, analysis of research, and the design of experiments. 2) Students of the new program achieved considerably more than did their controls in open-ended laboratory experiments. 3) Students in the new and the traditional programs did not differ significantly in their understanding of science (as measured by the TOUS - Test on Understanding Science), or in their general attitude toward biology. But students in the new program performed better than their controls on the SPI (Weich Science Process Inventory). 4) The new program was found to be better adapted to city and kibbutz schools than to agricultural schools. In fact, students of similar IQ who studied in agricultural schools achieved consistently less than equally able students in city and kibbutz schools. 5) The new program was found inadequate for a large portion of the lower IQ students whose parents had emigrated from Asian and North African countries. 6) Performance in biology was found to be multidimensional. By that we mean that different measures, such as paper-and-pencil tests and practical laboratory tests, yielded different results. A given student might reach different levels of achievement on different measures. Hence a comprehensive and valid evaluation would require a variety of measures, representing different formats of tests as well as different types of tasks (a practice not employed in the U.S.). 7) External examinations¹ may be profitably used to direct the implementation of a new program and promote practices congruent with those of the program's designers. 8) In principle, it is possible to obtain respectable achievements while studying a curriculum designed in one country and adapted by another, provided that local constraints receive the attention and treatment they deserve.

The evaluation in the first five years was designed to obtain information regarding the achievements and attitudes of the different types of students who studied the new program, thereby giving teachers and schools firm data as a basis for their decisions regarding adoption of the new curriculum. Yet, as with any research, the findings raised questions. As a result, a second generation of studies emerged.²

It is often argued that evaluation studies have very little effect on the actual development and implementation of curricula, either because the results are published too late or because curriculum developers are reluctant to rely on the results and tend to follow their personal impressions instead. This was not the case in the Israeli project. Table 1 offers several examples of how evaluation results have been used in curriculum improvement.

Finding	Consequent Action
There was low interest in the study of botany (Tamir and Jungwirth 1974).	Integrating the study of topics common to both botany and zoology (e.g. regulation, reproduction).
Teachers tended to teach about the same organisms that they studied in college (Tamir, 1976b)	Attending to the balance of topics presented in college biology courses.
Students had trouble applying statistics in biology (Jungwirth and Dreyfus 1972).	Preparing modules to suit the needs and interests of agricultural students.
Culturally deprived students whose families came from Asian and North African countries achieved poorly (Tamir and Jungwirth 1972).	Preparing a Hebrew version of the BSCS slow learners' test and designing supplementary modules; in-service courses for teachers.
Students completing grade 10 achieved poorly in inquiry skills; by the end of grade 12, most students reached a satisfactory level (Jungwirth 1970; Tamir and Jungwirth 1975).	Preparing instructional materials to promote inquiry skills, reducing the expectations for grade 10 students, allowing inquiry skills to develop gradually. Promoting use of <u>Invitations to Inquiry</u> .
Students showed poor understanding of some aspects of the nature of science (Jungwirth 1972).	Informing teachers of this finding; incorporating the history and philosophy of science in teacher training.
It was impossible to cover all the topics in "Diversity" in grade 10.	Designating some topics as core and the rest as electives.
Students had more trouble interpreting tables than graphs, and distinguished poorly between descriptions and interpretations (Jungwirth and Dreyfus 1972; Tamir and Jungwirth 1975).	Including in the matriculation examinations some problem situations, presented in tables, that required distinguishing between descriptions and interpretations.
Achievement scores on different tasks and from different test formats showed low correlations (Tamir 1974a).	Varying the tasks and test formats that constituted the matriculation examination, and urging teachers to use similar variety in their classroom evaluation.
Students of the new program achieved better than controls both on inquiry tasks and on knowledge of content (Jungwirth 1970, 1971; Tamir 1975a).	Acceptance of the new program by content-oriented teachers, and of the new matriculation examination as the national standard for biology majors.
Teachers used the matriculation examinations to guide their teaching (Tamir and Jungwirth 1975; Ben-Gal 1973).	Allocating funds for designing a better matriculation examination.

Table 1. The Effect of Evaluation Findings on the Development and Implementation of the Biology Program.

It may be useful to discuss some of the consequent actions in more detail. The finding that students in agricultural schools as well as those whose parents had immigrated to Israel from Asian and North African countries performed more poorly resulted in the establishment of three new curriculum development teams. One team developed special modules for agricultural schools which included topics of special interest to these schools, such as the mineral requirements of plants or the special nutritional problems of farm animals. In addition, many more examples of the application of biological principles in agriculture were included. Moreover, a special biology matriculation examination which addresses various agricultural aspects and applications was designed to match the objectives and subject matter of the agriculturally-oriented biology curriculum.

As to the slow learner, two separate teams developed two different curricula, one aimed at the lowest 15% of the school population and the other at the intermediate level, namely those who are below average but above the lowest 15%. High school biology teachers are now in a position to select from a variety of instructional materials which differ in their level of sophistication, in their cognitive demands and even in specific topics. Yet, in all of them, students are expected to be actively involved in investigations in the laboratory, in outdoor studies and in problem solving.

A special feature of the Israeli biology curriculum is the emphasis on inquiry. Various curriculum materials such as inquiry oriented laboratory investigations, invitations to inquiry, single topic films and television lessons have been developed or adapted from abroad. It is now possible for a teacher to use these materials to gradually develop inquiry skills in the students without pushing too hard too soon. For example, a student at the end of the 10th grade may not be required to design his own experiment but at the 11th and 12th grades students are expected to design their own experiments and to actually work according to their own design.

The Clientele

Lewy (1980) points out that the evaluation unit of the ICC aims mainly at meeting the needs of the project development teams. In the fifteen years of its existence it has rarely attempted to go beyond the early development stages. Consequently it has not engaged in the evaluation of the large scale implementation of programs. Naturally the evaluation reports, whether informal or formal, were written so that they would be able to serve the needs of the curriculum development teams. Nevertheless Lewy (1980) recommends that

- 1) Evaluation results should be presented to multiple clientele.
- 2) Evaluation results should be presented in a form which enables access of clientele not only to recommendations but also to the data on which the recommendations are based.
- 3) In spite of the fact that the preparation of formal written evaluation reports is time-consuming, there are several reasons to justify the preparation of such formal reports:
 - a) Orally or informally communicated findings are less likely to affect action, because it is easier for the decision-makers to disregard them.
 - b) The writing of a formal report will act as a safeguard for the quality and standards of the evaluation.
 - c) A sound theory of curriculum evaluation can be developed only on the basis of critical analysis of well written empirical studies.

The IHBP evaluation results have been communicated to a variety of audiences. Local reports were aimed primarily at teachers, administrators and curriculum developers. In addition a great number of studies were published in the professional educational literature in countries such as the U.S.A., U.K., Canada and Australia. In Israel, evaluation reports were published in teacher journals (*Maalot*), nature journals (*Teva Vaaretz*), curriculum literature (Theory and Practice, The Curriculum Center, the Ministry of Education and Culture), methods literature (*Methodica*, Tel Aviv University Publications) and quite frequently in the popular *Biology Teacher's Bulletin*, which has been published six to eight times yearly since 1968. The aim of this bulletin has been to present current information about various biological topics, to offer practical suggestions and classroom ideas, to provide teachers with a vehicle for presenting their own ideas, criticism and queries, and to present the results of studies and surveys concerning the development and implementation of the IHBP.

Following is a sample of titles related to curriculum evaluation which were published in the *Biology Teacher's Bulletin* in the years 1968-1980:

- Suggested changes in the matriculation examination
- How suitable is the syllabus at the 3 point level?
- Assuming individual projects
- Participation of teachers in designing the matriculation examination
- The role and structure of the oral examination
- The importance of evaluation for teachers

- A good word for the anathemic doctrine: teach for the test
- Examiners: examine yourselves
- The oral examination as a means for individualizing instruction
- The relationship between attitudes and nature field-work
- Assessing attitudes of students to environmental problems and to the simulation game, "Save our Environment"
- How students view their high school biology course
- Assessment of individual projects by the teachers
- How should one approach curriculum evaluation?
- A test based on single topic inquiry loop films
- The attitudes of Israeli biology teachers to the BSCS program
- Observations of instruction in high school biology laboratories
- Instructional activities in high school biology classes
- Botany and zoology in learning biology
- The views of teachers on the biology matriculation
- Analysis of the level of teachers' classroom and test questions
- Statistical information regarding the matriculation results
- What biology books and other printed materials do students use in their last two years of high school?
- Studying the Israeli BSCS program - problems and achievements
- A critical review of the book *How to Raise Animals*
- Quantitative analysis of commonly used high school biology texts
- Assessment of the instructional use of a slide series dealing with ecology
- Assessment of loop films dealing with human reproduction
- Assessment of individualized instruction of a unit on cell structure

Formal studies have not been conducted to ascertain the impact of these studies on teachers and administrators, but the popularity of the journal and informal responses and comments of teachers show that at the very least they read the findings.

Since 1970 about 50 evaluation studies related to IHBP have been published, in the U.S.A. (24), U.K. (16) and other countries (10). Of these about half deal with students' achievements, a quarter with implementation and classroom utilization and the rest with teachers' perceptions and behaviors. That these studies were welcomed by close to twenty respectable journals all over the world³ may be partly a result of the fact that the BSCS has been widely used in the U.S.A. and in many other countries. At the same time, however, it may be an indication of the wide interest in these evaluation studies, as well,

perhaps, as their cumulative nature, something rarely found in other places.

Evaluation papers related to IHBP have been presented in a number of professional conferences organized by associations such as the Asian Association of Biological Education, the American Association of Research in Science Teaching, the American Educational Research Association, the Australian Science Educational Association, the Israel Educational Research Association, the Bat Sheva International Conference and the International Conference of the Atlantic Institute of Education. Thus, the impact of evaluation of the IHBP has been significant and broad. Several evaluation instruments which were developed under the framework of the IHBP have been used in other countries (e.g. inquiry-oriented laboratory tests, cognitive preference inventories). As in other curriculum development projects located at universities (e.g. Individualized Science at the University of Pittsburgh, Project Physics at Harvard, and the Humanities Project at the University of East Anglia) a number of evaluation studies led to more general research dealing with problems such as misconceptions, logical errors, quantitative understanding, modes of performance, cognitive preferences, attitudes of students toward the use of animals in learning biology, the interests of students in different biological topics, the role of the laboratory in instruction and the effect of the school environment on students' learning.

Recently a new evaluation strategy - retrospective curriculum evaluation - was developed for assessing the long-term effects of the IHBP, which has shown, for example, that 11% of the variance in students' achievement in biology in their first year at the university can be accounted for by the nature of their high school curriculum (Tamir, 1980).

Lastly, an important feature of evaluation operating as part of a long-term curriculum project is the establishment of permanent facilities such as a test item bank; standard computer programs; a pre-arranged sample plan, with indices of stratification; existence of a full list of subjects included in the target population; existence of contacts and liaison persons in schools; different departments of the Office of Education, universities and other interested institutions; as well as advanced planning of computer analysis. During the years the entire project team became involved in various aspects of evaluation through the involvement of the team in the construction of the matriculation examinations which provide feedback on a regular basis.

THE IMPACT OF ASSESSMENT ON IMPLEMENTATION

Assessment and evaluation are extremely powerful means of affecting what is taught in schools and how it is taught. Following are examples of such effects brought about by the evaluation of the IHBP, especially by the changes in the matriculation examinations.

Examinations in schools. In most schools assessment has been based on written examinations. These examinations have changed considerably to keep pace with the adoption of new curricula in the past ten years. The impact of the external matriculation examinations, as described above, has stimulated teachers to design examinations that reflect a variety of objectives (i.e., various levels of Bloom's taxonomy, or inquiry objectives such as problem identification, hypothesis formulation, experimental design, and statistical analysis of data), as well as to use several test formats including multiple choice examinations, inferences drawn from data presented in tables and graphs, short open questions, and the analysis of research papers (Tamir, 1977).

The Israel High School Biology Project has made special efforts to publish and thereby make available to teachers different types of test items. Test design and analysis of results have been continuously included in inservice training courses.

In addition to improved written tests, practical examinations have been emphasized in secondary schools. Here again the three types of practical tests used in the matriculation examination have been adopted by school teachers as part of their own assessment schemes.

The involvement of teachers. The involvement of teachers take place in a number of ways: some teachers participate in the actual design and development of laboratory and paper-and-pencil tests. Many teachers serve as examiners for the different parts of the practical test: teachers from one school are assigned to test students in other schools. Not only do these teachers benefit from the guidance sessions which are held each year in preparation for the examinations, but, significantly, they gain tremendously by interacting with students who study the same curriculum from other teachers in different settings. The grading of the different written tests is done by teachers under close supervision - another golden opportunity for the self-development of teachers. Lastly, as described above, teachers always make special efforts to participate in inservice training days devoted to the previous year's matriculation examinations. The discussions on these occasions often stimulate second thoughts

among teachers as to what they have been doing and what can be done in their teaching. These occasions contribute considerably to the socialization of novice teachers. They also provide an opportunity for misunderstandings to be clarified and allow for fruitful exchanges between teachers and the project's staff.

Equipment and learning media. In two studies (Tamir, 1978a; Dreyfus, 1979), it was shown that as a result of the laboratory examinations, not only are students performing laboratory investigations regularly, but, almost with no exception, schools have acquired well-equipped laboratories which improve yearly in terms of materials and facilities. Also, most schools employ laboratory technicians who facilitate learning in the laboratory.

Learning materials and experiences. It has already been mentioned that the tests themselves become valuable learning materials. However, the impact of the examinations goes beyond that. For example, the fact that the written examination contains a piece of research stimulates many teachers to direct their students to read research papers. Another example: the need to learn to use an identification key results in field trips and incidental learning about different plants in their natural habitats.

Attracting students. Ten years ago biology was a relatively unattractive topic in high school, usually avoided by the more talented students, who elected to major in physics or chemistry. Today it is considered as a high level, highly respected, science course. About 50% of candidates admitted to medical school at the university have majored in high school biology.

Impact on college. University professors have begun to realize that students majoring in high school biology are better prepared for studies in the university. First year courses are now taking into account the preparation of students in high school biology as evidenced by their matriculation examination (Tamir, 1978b).

Teacher training exercises. A unique feature is the teacher training exercises based on the answers of examinees in the matriculation examinations. For example, the answers of students to the inquiry-oriented laboratory investigations were collected and used for exercises, each dealing with one of the following inquiry skills:

- Problem identification and formulation
- Hypothesis formulation
- Identifying the dependent variable
- Identifying the independent variables
- Identifying measurement procedures
- Identifying controls
- Designing experiments
- Designing controls
- Performing observations
- Performing measurements
- Reporting
- Designing graphs
- Designing tables
- Interpreting data
- Concluding
- Explaining
- Applying findings in a new context.

The fact that these exercises are based on actual answers of students, correct and incorrect, routine and creative, usual and unexpected, make these unique exercises an important means of helping teachers identify the deeper level of their students' learning, achievements and understandings. Some teachers have been using these exercises with their students and have found them highly profitable and rewarding.

In conclusion, it is suggested that a continuously operated curriculum project, located at an institute of higher education and directed by university personnel, may be a highly supportive setting for useful and productive curriculum evaluation.

NOTES

1. A brief description of the matriculation examination follows:
A theoretical test (paper-and-pencil) accounts for 60% of the total examination score, and is structured as follows:
Part A: Thirty to thirty-five multiple-choice items test for functional knowledge and comprehension.
Part B: Several problem situations are presented either verbally or, more frequently, in a table or graph. Students are required to describe the phenomena represented by the data, and to make inferences based on the data as well as on their ability to apply general biological principles.

Part C: Students are required to analyze an unfamiliar research paper. On the basis of that analysis, they must identify a new research question, formulate a hypothesis, and design an experiment to test the hypothesis. Each year a different research study is selected. A variety of topics has been included, such as growth and development, bird migration, hormonal regulation, the social behavior of bees, and the treatment of fruit after the harvest. For more details on the theoretical test, see Jungwirth and Dreyfus (1972).

A practical test accounts for 40 percent of the final score, and consists of three parts:

Part A: Students identify an unknown plant and an unknown animal with the aid of a key. They are required not only to identify the family, genus, and species, but also to record in sequence the characteristics that warrant their conclusions.

Part B: Students take an oral examination based on an ecological project which they have been performing for several months on topics of their own choice. Students bring their chosen organisms to the examination, where they must show a familiarity with the organisms and their environments. They must then use the organisms to demonstrate biological principles. The oral format permits the examiner to design a different test for each individual student. For more information, see Tamir (1972b).

Part C: A two-and-one-half-hour laboratory examination presents students with materials, organisms, and a problem which they must solve by performing an experiment, collecting data, making inferences, and properly discussing their findings. About ten new test problems have been designed each year. The examination can be administered to groups of ten to sixteen students simultaneously. Though the assessment is based mainly on the written report, some weight is also given to the students' use of materials and equipment. Thus the score is based on manipulation, experimental design, self-reliance, observation, communicative powers, and reasoning. For a fuller description of the laboratory examination, see Tamir and Glassman (1970, 1971a, 1971b), and Tamir (1974a).

2. Since the curriculum evaluation employed by various projects has usually stopped at the end of the first generation of studies, it may be of interest to list the questions that were selected for study. The following questions have been studied since 1971:
 - 1) To what extent are teachers aware of the program's objectives? What are their preferences and expectations regarding these objectives? (Tamir and Jungwirth 1972).
 - 2) What is the relationship between the teacher's awareness of objectives, the image of the teacher in the eyes of his students, and the students' achievement? (Jungwirth and Tamir 1973; Tamir 1975a).
 - 3) What is the relationship between teachers' attitudes toward the teaching of biology by inquiry and their adoption of the new curriculum? (Tamir 1976a).
 - 4) To what extent do student preferences for particular topics affect their achievement in these topics? (Tamir and Jungwirth 1974).
 - 5) What is the relationship of certain emphases in teacher training to the achievement of students in topics which have received differential emphases? (Tamir, 1976b).

- 6) To what extent has teaching the new curriculum changed teachers' attitudes? (Jungwirth 1975).
 - 7) What is the effect of the new program, as well as that of teachers' biases regarding the program, on the cognitive preferences of students? (Tamir 1975b).
 - 8) What is the relationship between achievement in biology and cognitive preferences? (Tamir 1975c).
 - 9) What accounts for the low achievement of students in agricultural schools? (Tamir 1972c; Dreyfus 1975).
 - 10) How does achievement in paper-and-pencil tests differ from achievement in practical laboratory tests? (Tamir 1972d).
 - 11) How do students and teachers regard various parts of the new matriculation examinations? (Tamir 1973).
 - 12) How has the new curriculum affected the number of laboratories, the equipment of laboratories, and the availability of assistance by laboratory technicians? (Tamir 1974c).
 - 13) What efficient teacher-training practices help prepare teachers to teach biology by inquiry? (in progress).
3. A list of journals in which evaluation studies of the IHBP have been published:

U.S.A. and Canada:

American Biology Teacher
 American Educational Research Journal
 BSCS Newsletter
 Curriculum Theory Network
 International Evaluation Newsletter
 Journal of College Science Teaching
 Journal of Educational Measurement
 Journal of Research in Science Teaching
 School Review
 Science Education
 School Science and Mathematics
 The Science Teacher

U.K.:

British Journal of Educational Psychology
 Journal of Biological Education
 Journal of Curriculum Studies
 School Science Review

Other countries:

Asian Association of Biological Education Newsletter
 Australian Science Teachers Journal
 European Journal of Science Education
 Studies in Educational Evaluation

REFERENCES

- BEN-GAL (GLASSMAN), S. A Good Word for the Anathemic Doctrine: Teach for the test! *Science Teacher*, 40(8), 5-6.
- BIOLOGICAL SCIENCES CURRICULUM STUDY *Biological Science: Molecules to Man*. Boston: Houghton Mifflin, 1963 (a).
- *High School Biology*. BSCS Green Version. Chicago: Rand McNally, 1963 (b).
- *Biological Sciences: An Inquiry into Life*. New York: Harcourt, Brace and World, 1963 (c).
- DREYFUS, A. Information Processing Within and Outside the School Curriculum by Students in Agricultural Schools as a Starting Point for Instruction. Ph.D. Thesis, Hebrew University of Jerusalem, 1975.
- DREYFUS, A. Evaluation of the Impact of the BSCS on the Laboratory Facilities in Schools. Unpublished paper, The Israel Science Teaching Center, 1979.
- GROBMAN, A.B. *The Changing Classroom*. Garden City, N.Y.: Doubleday, 1969.
- GROBMAN, Hulda G. *Developmental Curriculum Projects: Decision Points and Processes*. Itasca, Ill.: F.E. Peacock, 1970.
- JUNGWIRTH, E. Active Understanding of the Processes of Science. *Journal of Biological Education*, 1969, 3, 45-55.
- An Evaluation of the Attained Development of the Intellectual Skills Needed for "Understanding of the Nature of Scientific Enquiry" by BSCS pupils in Israel. *Journal of Research in Science Teaching*, 1970, 7, 141-51.
- Content-learning in a Process-Oriented Curriculum: Some Aspects of BSCS Biology in Israel. *Science Education*, 1971, 55, 85-96.
- TOUS Revisited: A Longitudinal Study of the Development of Understanding of Science. *Journal of Biological Education*, 1972, 6, 187-96.
- A Study of the Stability of Biology Teachers' Priorities and Expectations. *Journal of Curriculum Studies*, 1975, 7, 69-77.
- JUNGWIRTH, E. and DREYFUS, A. The Israeli "Bagrut" Examination in BSCS Biology. *Journal of Research in Science Teaching*, 1972, 9, 361-68.
- JUNGWIRTH, E. and TAMIR, P. The Teacher's Image as a Predictor of Students' Achievement. *Journal of Biological Education*, 1973, 7, 40-44.
- LEINHARDT, G. Outcomes of a Strategy for Program Evaluation. *Science Education*, 1977, 61, 1-17.
- LEWY, A. The Nature of Curriculum Evaluation. In A. Lewy (Ed.) *Handbook of Curriculum Evaluation*, UNESCO Paris & Longman New York, 1977, 3-36.
- Continuous Evaluation Services Attached to Large Scale Action Projects. Israel Curriculum Center, Ministry of Education and Culture, 1980 (mimeographed).
- NUFFIELD FOUNDATION SCIENCE TEACHING PROJECT *Synopsis of the Nuffield O-Level Biology Course*. London: Nuffield Foundation, 1965 (a).
- *A-Level Biology: Aims and Outline Scheme*. London: Nuffield Foundation, 1965 (b).

- TAMIR, P. Understanding the Processes of Science by Students Exposed to Different Science Curricula in Israel. *Journal of Research in Science Teaching*, 1972, 9, 239-45 (a).
- _____ The Role of the Oral Examination in Biology. *School Science Review*, 1972, 54, 162-65 (b).
- _____ Biology and Agriculture: A Problem in Education. *Journal of College Science Teaching*, 1972, 1, 48-50 (c).
- _____ The Practical Mode - A Distinct Mode of Performance in Biology. *Journal of Biological Education*, 1972, 6, 175-82 (d).
- _____ Attitudes of Students and Teachers Towards the Practical Examination in Biology. *BSCS Newsletter*, 1973, 53, 2-5.
- _____ An Inquiry Oriented Laboratory Examination. *Journal of Educational Measurements*, 1974, 11, 25-33 (a).
- _____ Attitudes of High School Biology Teachers to the BSCS Program in Israel. *Studies in Educational Evaluation*, 1976, 2, 123-129 (a).
- _____ The Impact of the BSCS on Laboratory Facilities for Biology Students in Israeli High Schools. *International Evaluation Newsletter*, 1978, 19, 20-23 (a).
- _____ The Effect of the Teachers' Curricular Bias, the Kind of Curriculum Studied, as well as their Interaction on Students' Achievement in High School Biology. Paper presented at the forty-eighth annual meeting of the National Association of Research in Science Teaching, Los Angeles, March 1975 (a).
- _____ The Relationship between Cognitive Preference Subject Matter. *American Educational Research Journal*, 1975, 12, 235-264 (b).
- _____ The Teaching and Achievement in Botany and Zoology as Affected by the Teachers' Background. *Science Education*, 1976, 60, 291-298 (b).
- _____ Questioning Practices in the Teaching of High School Biology in Israel. *Journal of Curriculum Studies*, 1977, 9, 145-156.
- _____ The Potential Role of Examination in Innovative Curricula. *The American Biology Teacher*, 1978, 40, 353-357 (b).
- _____ Retrospective Curriculum Evaluation - An Approach to Longterm Curriculum Evaluation, Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980.
- TAMIR, P. and GLASSMAN, F. A Practical Examination for BSCS Students. *Journal of Research in Science Teaching*, 1971, 8, 307-15 (a).
- _____ A Practical Examination for BSCS Students: A Progress Report. *Journal of Research in Science Teaching*, 1971, 8, 307-15 (a).
- _____ Laboratory Test for BSCS Students. *BSCS Newsletter*, 1971, 42, 9-13 (b).
- TAMIR, P. and JUNGWIRTH, E. Teaching Objectives in Biology, Priorities and Expectations. *Science Education*, 1972, 56, 31-39.
- _____ Students' Growth Developed as a Result of Studying BSCS Biology for Several Years. *Journal of Research in Science Teaching*, 1975, 12, 263-80.
- _____ Botany and Zoology - A Curriculum Problem. *Journal of Research in Science Teaching*, 1974, 11, 5-16.
- WELCH, Wayne W. The Impact of the National Curriculum Projects: The Need for Accurate Assessment. *School Science and Mathematics*, 1968, 68, 225-34.
- WELCH, W.W., and WALBERG, H.J. A National Experiment in Curriculum Evaluation. *American Educational Research Journal*, 1972, 9, 373-383.

STANDARDIZED EVALUATION OF EDUCATIONAL PROGRAMS

Daniel J. Davis

In Israel there is serious scepticism among educational researchers whether it is possible to carry out meaningful evaluation of educational programs in schools. The reason for this scepticism is that educational evaluations are field studies and suffer from all the handicaps associated with that kind of research. In addition to this, the educational evaluator must contend with disturbing factors which make failure almost certain:

1. In many cases it is not clear what the program is intended to do.
2. In many cases it is not clear how the program is to operate.
3. In most cases it is not known whether, in fact, the program was actually implemented in particular locations; and if it was implemented, whether the implementation was adequate. (This can only be answered if it is clear how the program is to operate).
4. In almost all cases there are no tailored instruments for measuring the outcomes, so the evaluator is forced to develop his own or to use standard instruments. Developing tests, questionnaires, interview schedules, or observation schedules is very difficult, expensive and time-consuming. Using a standard achievement test means that the researcher is in many cases actually investigating transfer of learning before he has checked for learning. This is most problematic when the program is focused and the test highly general.
5. In most cases the evaluator comes on the scene after the program has been initiated. In this case he will generally search out an appropriate control group. The control group will generally be matched with the experimental group on socio-economic status, which is clearly inadequate for correcting for selection.

Even pretest data is problematic since selection into a program is often based on motivation of children and parents, etc., factors which are

generally not controlled.

6. In most cases the program is tested against a "no-treatment" control. In fact, the controls may be getting equivalent or alternative treatments from other sources. Ironically, the presence of the treatment may have led to the channeling of other resources to the "no-treatment" group.
7. The demands of the evaluation are necessarily secondary to the demands of the program and to the demands of the school. Therefore, attempts by the evaluator to obtain important information are likely to be frustrated when they conflict with school and program demands.
8. In many cases the program implementer does not particularly want the evaluation but has been forced to accept it for political reasons. Since his full cooperation is essential this factor alone can prevent the success of the evaluation.
9. Since educational processes tend to be slow and children's development erratic, it may take a long time for the effects of a program to be manifested. Since there is generally pressure for quick results, in many cases the evaluator is forced to examine these too early and to rely on too few measurements.
10. Since the term of an educational project is generally long (a year or more) evaluation is particularly vulnerable to dropout.

What can be done in the light of such overwhelming odds for failure? The recommendation presented here is that educational evaluation as it is conducted today be dropped and be replaced by an evaluation strategy which eliminates or reduces the uncertainties created by the above factors. Such a strategy focuses its efforts on programs with relatively clear-cut goals and operating procedures (factors 1 and 2). Instead of asking how well the program works in general, it first asks how well the program can work, and attempts to answer this question in a tight experimental framework. The result can be viewed as a standard against which future evaluation results in normal field settings can be compared.

The following discussion outlines the procedures associated with this strategy, which is called standardized evaluation. This is contrasted with generalized evaluation which is conducted in the normal field setting of a program.

A standardized evaluation is one in which two conditions are present.

1. The program is implemented in its "standard form" - the field conditions, the personnel and operation are made to approximate optimization according to the conception underlying the program.
2. The evaluation conditions are set up to approximate the ideals of experimental research - the evaluator has control of selection to treatments; the process of program implementation is closely monitored; and outcomes are carefully and repeatedly measured.

The purposes of the two conditions are: 1) to make clear what is being evaluated; 2) to make sure that the "what" is an exemplary implementation of the program; that is, a reasonable approximation of the optimization of program parameters according to the conception underlying the program; 3) to make sure that the estimates of program effects are as accurate as possible.

I. THE STANDARD PROGRAM

Under the assumption that the program parameters have been optimized, the resulting estimates of program effects can be viewed as upper bounds to the effects of program implementation in normal field conditions. That is, they provide standards against which normal field implementation of the program can be evaluated.

IA. An example - Individualized Tutoring and Support of Disadvantaged Children

In Israel there is a strong emphasis on individualized tutoring programs for improving the social and cognitive functioning of disadvantaged children. For example, approximately 4000 university students throughout Israel tutor children in grades 5 through 9, four hours per week in two sessions. Also, about ten teachers colleges are participating in a program in which students tutor elementary school children under the guidance of a member of the staff. Plans are in progress for widening this program so that all teachers college students will tutor a disadvantaged child at some point in their training periods. In addition to these, there are a large number of programs utilizing college and high school students as tutors.

Several evaluation studies have shown rather weak but positive effects of the tutoring on educational achievement (e.g. Sprinzak, 1975) and others have shown weak negative effects (Prinz, 1977; Eisenberg, *et al.*, 1980). Since these evaluations suffered from the difficulties listed at the start of this paper I argue that the results are highly questionable. In particular, it is not clear whether the tutoring took place as it should have, nor is it clear

whether when difficulties arose the tutors obtained support and guidance from experienced supervisors. Therefore, the finding of "weak effect" (negative or positive) may be as much a result of a lack of treatment as it is of the effect of treatment when applied.

After a review of the research literature, a standard program in the area of individualized tutoring and support of disadvantaged children was specified (Davis & Snapirit, 1979; Davis, Snapirit & Golan, 1980). The main feature of the program is a supervisory and guidance structure which is sensitive to the problems encountered by the tutors and can help in solving them. Also, the program is maximally flexible so that it can adapt to the specific conditions of each tutor-tutee relationship.

The framework for the program is an undergraduate course in the School of Education at the Hebrew University in which the students receive half scholarships in addition to credit for the course. Each student works with one child in grades four to five at least twice a week for two hours each time. Each tutor receives an hour per week of individual guidance with an experienced supervisor in the field of individual tutoring. The class framework is used for group discussions and lectures. As part of the course requirements, each student prepares detailed descriptions of the meetings with the child.

As a result of the present pilot operation of the program, we feel that it represents a reasonable approximation of an optimal application of an individualized tutoring program. Therefore the estimates of its effects may serve as standards for evaluating other programs. In fact, in a research design (described later) the above program is being compared with a field program in which there is minimal supervision and guidance. This is the preferred approach to comparing the standard program with field programs. However, it will generally not be possible to do this because of clashes between the demands of research of a standardized evaluation and those of field programs.

IB. Interpreting the Results of Standardized Evaluations

Case 1: The standardized evaluation results in an estimate of no effect or negligible effects. One could legitimately question the efficacy of the program in any form and recommend either dropping or revising it.

Case 2: The standardized evaluation results in a large program effect and all consequent generalized evaluations result in negligible effects.

In this case the program director must consider alternative ways for improving the field applications of the program. If none are found, he must consider implementing the program in its standard form or dropping it as too expensive.

Case 3: The standardized evaluation results in a large program effect and at least one of the generalized evaluations results in a moderate effect. In this case the program director must consider the importance of the difference in effect between the standard and the field program in relation to the difference in their costs. On the basis of a cost-benefit analysis he can decide whether or not to revise the field program.

Case 4: The standardized and generalized evaluations result in approximately the same moderate or large program effects. In this case it can be argued that the program effects are robust with respect to changes from standard program conditions to normal field conditions in the implementation of the program. The program director is in the enviable position of stating that at least one of the normal field implementations has produced maximal results according to the conception underlying the program. He can then consider ways for reducing the cost of the program without reducing its effectiveness.

Each of the above cases provides valuable information for the implementation and evaluation of educational programs. This is a result of the provision of a standard for evaluating program effects.

It is worth emphasizing that the standard program is assumed to be optimal according to the conceptual framework. If in an actual study a field program yields better results than the standard program, then the conceptual framework which led to the development of the standard program and to the development of the field program must be called into question.

IC. Criticisms and Responses

Three criticisms are generally brought forth concerning the conceptual framework of standardized evaluation as well as its utility. I will relate to these here:

Criticism 1:

"How can we claim to have created an optimal program when in fact we have only vague theoretical notions and the skimpiest empirical evidence about factors which do or do not affect the program's effectiveness? Instead,

should we not start with a factorial design in which the hypothesized important variables are manipulated and the various combinations examined? Only then can the conception of an optimal program be used."

My first response to this is that in many educational programs there is enough specification of desired results and operating procedures to allow for general agreement on a standard program within the conceptual framework underlying the program. In fact, it can be claimed that such specification is a prerequisite to meaningful evaluation. When there is no such specification it would seem that in many cases intensive discussion and consultation with experts would lead to agreement on a standard program.

Clearly, it is possible that more than one standard program will be specified. For example, in the case of our research on individualized tutoring, we are considering two different administrative frameworks for providing guidance and support to tutors. Our expectation is that they will result in similar effects on achievement. However, if this does not turn out to be the case, the better program will be used as a standard.

My second response to the criticism is that the factorial design represents a form of fine tuning which should be attempted only after one has obtained relatively gross indications that the program has some effect in one or more of its possible forms. In addition, factorial experiments are impractical in most educational settings where more than two or three alternative treatments in a reasonable experimental design lead to huge costs.

I am arguing that educational theory and research at this point in time require the answering of relatively gross questions such as "Can this program work under the best conditions considered over the combination of all program parameters?" This is a gross question since after the study we will have no formal evidence concerning which aspects of the program were important and which were not important. To some this may seem inadequate against the richness of possible conclusions to be obtained from and the exactness promised by the classical designs of psychology and educational psychology. The point being made here is that much of the malaise in educational evaluation (at least in Israel) is a result of skipping over the stage of answering the gross question (How well does the program work when implemented under optimal conditions?).

Further, it is my contention that educational research will probably never become the fine art of tuning that is implied in many factorial studies. The multiplicity of interacting factors precludes exact specification. The best

we can do is to specify that certain combinations of factors are important or that two programs which are different on a number of dimensions led to the difference. To speak of changes in the effect of a program in terms of changes on a single dimension while holding all other dimensions constant ignores the complex interdependencies among educational inputs, and will most likely lead to misleading or meaningless results. However, even if this contention is not accepted, the argument holds that the fine tuning stage should be preceded by an estimate of the effect of the program in a standardized evaluation.

Criticism 2:

"Even if a standard program could be specified, demonstrating its effectiveness is trivial". This is based on the common assumption that optimization of almost all reasonably focused educational programs will lead to important changes in educational performance. This is a manifestation of our never ending optimism that any well-intended action will have some beneficial effect no matter how poorly conceived. I see no justification for this optimism and even if it were justified I do not agree that it makes standardized evaluation trivial. Even if we were certain that individualized tutoring would lead to improved school performance when provided under optimal conditions, it would still be important to estimate the size of the improvement. How much of an improvement can we get when optimal individualized tutoring is provided for children with learning difficulties? This, I claim, we do not know, and it is worth knowing. It will provide a sound standard for judging the effectiveness of actual field programs and will provide useful information for improvement.

In fact, Riecken and Boruch suggest that social experiments include "an extremely intensive treatment in order to learn whether the treatment is at all effective under the most favorable circumstances" (p. 50, 1979). This approach is reflected in a study by McKay, *et al* (1978) which examines the effects of a combined nutritional, health care, and educational program on cognitive ability.

Criticism 3:

"Even if the standard program were determined and its effect accurately estimated as being large it would certainly be impractical. It would require a tremendous investment of resources if applied on a large scale." This criticism reflects a basic misunderstanding of the concept of standardized evaluation. The purpose of the standard program is to provide a standard

for evaluating the effectiveness of field programs and not necessarily to provide that the educational decision makers will decide that the standard program is worth implementing on a wide scale due to its greater effectiveness. A possible conclusion based on standardized evaluation is that educational programs must be more intensive in order to lead to meaningful and lasting changes. However, even if this did not turn out to be the case, standardized evaluation would still provide valuable information for evaluating and improving the normal field implementations of educational programs.

II. EXPERIMENTAL DESIGN

The second feature of standardized evaluation is that the evaluation design is made to approximate laboratory research models. In most cases of field programs this is virtually impossible as a result of the factors listed at the beginning of this paper. While I accept Boruch and Rindskopf's (1977) claim that randomized field studies are feasible in many cases, there are so many factors working against their success in normal field settings that serious measures are necessary to insure satisfactory experimental conditions. Two measures suggested here are an integral part of standardized evaluation:

1. The overall project (including both implementation and evaluation) is under the control of the evaluator. That is, the program director in schools participating in the project is subordinate to the evaluator.
2. Selection of participating schools and classes is based on the presence of a strong need and desire for the program.

IIA. The Evaluator as the Key Person

In the natural order of events in the field, program considerations will usually dominate research considerations. This is as it should be. However, Boruch and Rindskopf (1977) contend that it is possible to blend research and program considerations so that both sets of demands are satisfied. Unfortunately, program directors are not likely to be sensitive to research considerations and the desired blend will be very difficult to achieve. On the other hand, the evaluator is very likely to be sensitive to program demands in addition to research demands. In fact, it is essential in general that the evaluator be sensitive to program demands in order for successful evaluation to occur. Therefore, it is argued here that the most harmonious blend of jointly satisfying program and evaluation considerations will be

obtained when the evaluator has ultimate responsibility for both the program and the evaluation. This does not mean that the evaluator runs the program directly. Rather, it means that the program director is forced to seriously consider research demands along with the evaluator. In other words the evaluator does not come after the fact or as a somewhat annoying appendage to the research project, but is an integral and key part of the overall research program (implementation and evaluation).

One could argue that the program will be "spoiled" by the "intrusion" of the evaluator. This is certainly a possibility that must be considered, and efforts must be made to prevent its occurrence. The assumption underlying this approach, however, is that the evaluator is sensitive to the subtle interplay of program-related and evaluation-related decisions. If this is the case, then it is not likely that either the program or the recipients will be harmed by evaluation considerations in program decision-making.

IIB. Research as a Contributor to School Welfare

The second measure concerns the manner in which schools are approached to participate in the research (program and evaluation). To the extent possible the program should be something which is viewed by school personnel as a desirable addition to or a replacement of a present program. Thus, the research can be seen as something which will contribute to the school's welfare, and the demand of the research for meeting design requirements is then more likely to be accepted by the school. Also, the fact that the evaluator is the key representative of the research to the school greatly increases chances that evaluation considerations will be approached seriously by the school personnel.

My experience indicates that this approach will generally lead to complete cooperation of the school on four key research issues:

1. selection to treatment;
2. monitoring of program implementation;
3. measurement of program outcomes;
4. control and monitoring of alternative treatments.

For example, selection of children into experimental and control groups is the most sensitive topic in which school and research demands clash. The school wants all those who need it to participate in the program. The research wants the program candidates to be divided randomly or by a well specified rule plus randomization into two groups, one of which will be

deprived of the program. In my experience, careful explanation before creating the pool of candidates, along with emphasis on the fact that accepting the research will increase the number of children participating in the program, has led to the school's acceptance of the evaluation demands. This strongly refutes the claim by many educational evaluators in Israel that randomization is impossible in the school setting.

While selection for treatment is the most immediate problem when discussing the implementation of the program and the evaluation in the school it should not overshadow other serious issues.

An important feature of standardized evaluation is close monitoring of program implementation. This provides the evaluation and operation teams with information on the extent of treatment application in the sample. If it is inadequate, measures can be taken to improve it. If it is adequate, the information can be used to provide a measure of implementation efficiency in normal field applications of the program. Close monitoring is dependent on the cooperation of school personnel. Therefore, discussions with school personnel prior to agreement on cooperative effort should emphasize the research demands for monitoring. The relevant personnel should be the ones party to the agreement. In most cases it is the teacher who is the key and who should be consulted early in the discussions of implementation in the school.

In addition to playing a vital role in monitoring activities, the teacher is particularly important in obtaining outcome measures. The third key feature of the standardized evaluation is that outcome measures are repeatedly measured to insure stable estimates of program effects. In addition, tests are designed around the class curriculum in order to be maximally sensitive to direct program effects. (Boruch and Gomez have suggested a model of field evaluation efficiency in which a key element is the extent to which a test directly reflects program related behaviors, 1977). It is my opinion that the classical educational design of a single pretest and a single post-test using a standard achievement test is so crude in its measurement of program related behaviors and so susceptible to the vagaries of classroom conditions during testing that its power for detecting program effects is close to zero. The only way I can see for increasing the sensitivity of outcome measurement is through focused and repeated measurement. Again, need and desire for the program should greatly increase the teacher's willingness to cooperate on this aspect of the research. In addition,

some teachers have viewed the ongoing testing as a source of information which is a help to them.

The fourth issue, which is the most problematic, concerns the presence of a number of alternative treatment frameworks in a single school. The problem for the research is that selection of a child to either the control or experimental group may affect his or her selection to another treatment framework. In addition, it may be that after the implementation of the experiment in a school another program may be introduced into the school. Obviously the researcher cannot demand that children in the control group be deprived of alternative programs if they are available. However, he can reduce the effect of alternative programs by: 1) selecting schools which have relatively few treatment frameworks and which are not likely to introduce new programs during the period of the experiment; and 2) screening out children who are already in an alternative program or are being considered for one. Again, close cooperation of the school staff is required to deal with this complex issue and even close cooperation does not guarantee complete success.

IIC. Continuation of Individualized Tutoring Example

The standard tutoring program described above is being examined in a number of elementary schools within an experimental design based on the procedures discussed in the previous section. In addition, a normal field application of individualized tutoring is being examined in the same experimental framework. The comparison of the field and standard programs is particularly interesting since the former (known as "Perach", see Eisenberg *et al*) is based on a conception of minimal guidance and supervision of the tutors. As mentioned previously, the standard program is based on a strong guidance and support structure in which tutors receive individual and group guidance regularly. Thus, two extremes of a continuum are being compared in a single experimental framework. (It should be noted that the achievement of comparing the standard and the normal field program in a single experimental framework is due to the exceptionally close cooperation of the school staff and the program coordinators of Perach with the research team).

The experimental design recognizes the classroom as the key frame of reference for examining school achievement. Therefore, the classroom is the unit of analysis and each class is viewed as an independent replication of the experiment. Within each class the children are divided into four groups. Three of these include the pool of children who were candidates

for the program. These are the group of children in the standard program, those in the normal field program and a no-tutoring control. The fourth group consists of the remainder of the class. The selection of program candidates to one of three groups is under the control of the research staff as prearranged with the school staff. Close monitoring of the standard program is a natural by-product of the guidance framework and is being maintained at a satisfactory level. Only partial monitoring is being achieved in the Perach framework; i.e., information based on teachers' reports and on the reports of the Perach coordinators, but there is no direct contact with the tutors.

The cooperating teachers have agreed to repeated testing over the school year using tailored as well as standardized tests, and have provided the research staff with the results of their own tests and with their grades from the present year and from previous years. The dependent variable for all grades and tests is the percentile rank of the child relative to the entire class; that is, the percent of children with results equal or less than that of the child in question. This is a standardized score which allows comparison of tests and grades given at various times and measured on different scales. This is possibly less sensitive than a z-score standardization, but it has the advantage of being free of assumptions concerning equality of distances on all points of the scale of measurement. In addition, the meaning of the percentile score is more readily apparent.

A key element in the analysis is repeated measurement over a long time period. This is the only way we can see for dealing with instability in children's test performance. Thus, the basic analysis is a comparison of time series data for children studying in the same class over a period of years. The fact that the school context (e.g., classmates and teachers) is constant for the experimental and control children, makes this a powerful design for investigating the effects of individualized treatments such as the tutoring investigated here.

The research staff is examining in detail the participation of children in alternative programs. So far, one child in one of the tutoring programs was found to be in a special class for children with behavior problems. This occurred in spite of careful explanation and the good intentions on the part of the school staff. Any imbalance among the control and experimental groups in terms of participation in alternative programs will be accounted for in the analysis of the data.

Since comparisons are for children in the same class, follow-up in terms of future grades and class administered test results is relatively easy.

III. SUMMARY

Ten points were listed which make educational evaluation in the natural field setting especially difficult. A strategy called "Standardized Evaluation" was suggested in which:

1. Programs with relatively clear-cut goals and operating procedures are selected for evaluation.
2. An evaluation expert is responsible for a research framework which includes both operation and evaluation of the program.
3. The program is implemented in its standard form. That is, the field conditions, the personnel, and operation represent a reasonable approximation of optimal implementation according to the conception underlying the program.
4. Schools and classes are selected for participation in accordance with their willingness to comply with the demands of the research design.
5. The research design approximates laboratory models in terms of selection to treatment, monitoring of program implementation, repeated measurement of program outcomes, and control and monitoring of alternative treatments.
6. The results of the evaluation provide a standard for evaluating the program in normal field settings. In many cases formal evaluation in the normal field setting may be unnecessary since the policy maker may have enough information of the implementation efficiency of the program in the normal and standard versions to come up with cost-benefit estimates.

REFERENCES

- BORUCH, R.F., and RINDSKOPF, D. On Randomized Experiments, Approximations in Experiments, and Data Analysis, in L. Rutman (Ed.) *Evaluation Research Methods: A Basic Guide*. Beverly Hills, Cal., Sage, 1977, pp. 141-176.
- BORUCH, R.F. and GOMEZ, H. Sensitivity, Bias, and Theory in Impact Evaluations. *Professional Psychology*, 1977, 411-434.
- DAVIS, D., SNAPIRIT, T. and GOLAN, P. An Individualized Tutorial Program: Operation and Evaluation, 1980, Research Proposal, The NCJW Research Institute for Innovation in Education, Hebrew University, Jerusalem (Hebrew).

- EISENBERG, T., FRESKO, B. and CARMELI, M. A Tutorial Project for Disadvantaged Children: An Evaluation of the Perach Project. 1980, Weizmann Institute, Rehovot, Israel.
- MCKAY, H., SINISTERPA, L., MCKAY A., GOMEZ, H. and LLOREDA, P. Improving Cognitive Ability in Chronically Deprived Children. *Science*, 1978, 200, 270-278.
- PRINZ, D. A Tutoring Project of Children Teaching Children - The Effects on the Tutees. 1977, Master's Thesis in Psychology, Hebrew University, Jerusalem (Hebrew).
- RIECKEN, H.W., and BORUCH, R.F. (Eds.). *Social Experimentation: A Method for Planning and Evaluating Social Intervention*. New York, Academic Press, 1974.
- SHPRINZAK, D. An Evaluation Study of Tutorial Projects, 1976, Final Report. The NCJW Research Institute for Innovation in Education, School of Education, Hebrew University, Jerusalem (Hebrew).

THE EVALUATION OF A MULTI-DIMENSIONAL PROJECT

David Nevo

Many evaluation activities do not succeed in serving decision-makers because either (a) they collect information that is irrelevant to the decisions to be made, and/or (b) no appropriate channels of communication are established between the evaluator and the decision-maker.

This paper presents an evaluation model for a multi-dimensional project intended to develop new modes of interaction between an interdisciplinary team of university faculty members and local educational communities. For such a project, a comprehensive evaluation mechanism is required to respond to the different evaluation needs of the project members and other project-related parties.

The first part of this paper describes the major characteristics of the project, identifies its evaluation needs and considers the difficulties involved in responding to such needs. The second part proposes a conceptual framework of evaluation to serve as a basis for the development of an evaluation mechanism for the project.

The last part of the paper presents an evaluation model for the project and specifies the organizational structure of the evaluation mechanism, its research design and measurement tools, and the communication system between the evaluation mechanism and the rest of the project.

Although this evaluation model was developed for a specific project, it might also be useful for other multi-dimensional projects concerned with the development of a systematic evaluation mechanism.

THE EVALUATION CHARACTERISTICS OF THE HELPING SYSTEM PROJECT

The evaluation model presented in this article was developed in the framework of the Helping System Project (HSP). As part of the general effort of the Ministry of Education to improve the educational system in disadvantaged communities in Israel, a group of faculty members of the School of Education of the Tel Aviv University undertook the responsibility to assist three local educational communities in improving their education systems. Having mixed feelings about the direct usefulness of some traditional research studies where university professors interact with the educational field to obtain their research data, this project tried, during a period of two years, to develop new modes of interaction between the university faculty members and the local educational communities attempting to cope with a wide range of educational problems in these communities.

The major principles of the Helping System Project and the basic characteristics of its implementation have been published in various documents (Smilansky & Sanders, 1975; Kashti, 1976), and it is not the purpose of this article to analyze those documents or to summarize them. The purpose of this section is to derive from those documents the multi-dimensional nature of the project and point out characteristics which seem to be relevant to evaluation and should be considered in designing an evaluation plan for the HSP or similar projects. The characteristics of the HSP can be summarized as follows:

- a) The HSP was an experimental project intended to develop nonconventional ways of interaction between university professors and practitioners. Being new, the project needed summative evaluation to demonstrate its merit as a significant way to cope with major educational problems and to justify its existence within the educational system. Being new also meant that the HSP needed formative evaluation to assist its staff in the "trial and error" process of developing an innovative course of action.
- b) This was a project of institutional learning intended to guide individuals and organizations towards a systematic process of problem solving. This had to include a rational process of problem identification, goal setting, choosing among alternative ways of action and assessing their efficiency. Not only had evaluation to be an integral part of such a process, but it was also important that the project which advocated such a rational way of action should demonstrate by example the use of rational behavior based on evaluative information.

- c) The project attempted to cope with real problems in the daily life of educational communities. In most cases, the "real world" is not ready to force itself into the mould required by experimental design studies. Thus the evaluator has to develop evaluation designs that will create a balance between standards of scientific accuracy and practical significance. The creation of such a balance is one of the most difficult challenges for an evaluator.
- d) In addition to its attempt to assist local communities, the HSP also intended to develop a generalizable model for interaction between university faculty members and practitioners. Such a model would have to be supported by data that could be generalized beyond the specific setting of the communities participating in the HSP. This situation could present problems for the evaluator who had to provide information that would be both significant within a specific context and yet generalizable within more universal settings.
- e) The project was involved in a continuous process of interaction with local communities. It therefore needed an evaluation mechanism that would provide ongoing evaluative information rather than a "one shot" research design that would assess some specific aspects of the project.
- f) HSP was acting simultaneously in many areas of the local educational system and trying to introduce changes in many variables related to the school, the local community and its leadership. The evaluation had to consider the multi-dimensional nature of the project and to develop a comprehensive evaluation network that would not limit itself to a limited number of variables such as changes in students' achievements or the improvement of the curriculum.
- g) HSP had an interdisciplinary staff comprised of psychologists, sociologists, an anthropologist and some educators. The utilization of a whole range of expertise required appropriate channels of communication for the extensive interchange of ideas among staff members. An evaluation mechanism could facilitate the communication process by providing it with a solid data base and a continuous stream of information.
- h) The HSP staff was comprised of persons whose work on the project was not their major professional activity. Most of the staff members were university faculty members whose major activities were teaching and research. The limited amount of time that they could devote to the project had to be utilized efficiently by the project. Thus, the evaluation mechanism

of the project had to not only provide information on the extent to which the project was utilizing its human resources efficiently, but also had to provide staff members with various kinds of information which they needed but did not have the time to collect.

- i) HSP's budget was large when compared with the amount of resources allocated to similar projects in the Israeli educational system. Thus, it could be expected that considerable resources be allocated for the conduct of evaluation. An appropriation of about ten to twenty percent of HSP's budget could provide a solid basis for appropriate evaluation of the project.

The above mentioned characteristics of the HSP point out why evaluation was an integral part of the project and also identify some of the problems involved in the conduct of such an evaluation. A comprehensive perception of evaluation was required to serve as a basis for the development of an evaluation mechanism that would be able to respond to the evaluation needs of this project. Such a perception of evaluation is presented in the following section.

THE CONCEPTUAL FRAMEWORK FOR THE EVALUATION

Before presenting the evaluation model developed for the HSP, an attempt will be made to clarify the meaning of "evaluation" and to present the perception of evaluation which served as a basis for the development of the evaluation mechanism presented in this article.

In recent years, many writers (Tyler, 1950; Cronbach, 1963; Scriven, 1967, 1972; Stake, 1967; Stufflebeam, 1969, 1974) have dealt with the concept of evaluation and have addressed various aspects of this concept. Below are ten aspects of evaluation which the writer perceives as being the major dimensions in the conceptualization of evaluation. Various aspects of these dimensions will be discussed in a way that will introduce a perception of evaluation which will be the basis for the rest of this article. The ten dimensions in the conceptualization of evaluation¹ are: 1) the definition of evaluation, 2) the function of evaluation, 3) the objects of evaluation, 4) evaluation variables, 5) evaluation criteria, 6) evaluation audiences, 7) the process of evaluation, 8) evaluation methods of inquiry, 9) types of evaluators, and 10) evaluation standards. The dimensions represent ten questions that should be addressed in the development of any evaluation design. The questions are presented in Table 1, followed by a discussion of their general meaning and their specific significance for the evaluation of the HSP.

Table 1. Questions to be Addressed by a Conceptual Framework for Educational Evaluation.

1	Definition	What is evaluation?
2	Function	What is evaluation for?
3	Objects	What are the objects of evaluation?
4	Variables	What variables (regarding each object) should be investigated by evaluation?
5	Criteria	What criteria should be used to judge the merit of an evaluation object?
6	Audiences	Who should be served by evaluation?
7	Process	What is the process of performing evaluation?
8	Methods	What are the methods of inquiry to be used in evaluation?
9	Evaluators	Who should perform evaluation?
10	Standards	By what standards should evaluation be judged?

1) What is evaluation?

In recent years a considerable degree of consensus has been reached among evaluators regarding the definition of evaluation as the assessment of merit or worth (Scriven, 1967; Glass, 1969; Stufflebeam, 1974). A joint committee on standards for evaluation, comprised of seventeen members representing twelve organizations associated with educational evaluation, recently published its definition of evaluation as being "the systematic investigation of the worth or merit of some object" (Joint Committee, 1981, p. 12). Such a definition, which stresses the judgmental character of evaluation, might create considerable anxiety among potential evaluatees and raise resistance among opponents of evaluation. This might explain some attempts that have been made (Cronbach, 1963; Stufflebeam, *et al.*, 1971) to present evaluation as a process to provide information for decision-making which does not necessarily involve any judgemental act. However, it seems to be unrealistic to try to create positive attitudes toward evaluation by ignoring one of its major

features. Another approach intended to develop positive attitudes toward evaluation might be by demonstrating its constructive functions within the various domains of education (Stufflebeam, 1969; Stufflebeam *et al.*, 1971).

2) What is evaluation for?

Scriven (1967) was the first to suggest the distinction between "Formative Evaluation" and "Summative Evaluation" as two major roles or functions of evaluation, although he was probably not the first to realize the importance of such a distinction. Later on, referring to the same two functions, Stufflebeam (1971) suggested the distinction between proactive evaluation intended to serve decision-making, and retroactive evaluation intended to serve accountability. Formative evaluation is used for the improvement and development of an ongoing educational activity. On the other hand, summative evaluation is used for accountability, certification or selection. A third function of evaluation, which has been somehow ignored in the literature, should also be considered. This is the psychological or socio-political function of evaluation. In too many cases is it apparent that evaluation is not serving any formative purposes nor is it being used for accountability or other summative purposes. It is being used, however, to increase awareness of special activities, to motivate evaluatees, to achieve desired behavior, or to promote public relations. Regardless of our personal feelings regarding the use (or misuse) of evaluation for such purposes, we cannot ignore this additional function of the process.

3) What are the objects of evaluation?

Almost any educational entity can serve as an object of evaluation. While some objects, like students or school personnel, have always been popular objects of evaluation in education, some others, such as projects, programs curricular materials or educational institutions, have also become favorite objects of evaluation in recent years. The clear identification of the evaluation object is an important part in the development of any evaluation plan.

4) What variables should be evaluated by evaluation?

Regarding each evaluation object, a decision has to be made about the various aspects of the object that should be evaluated. Those various aspects are the evaluation variables. Traditional approaches to evaluation focused mainly on products or outcomes. Thus, to evaluate an educational object (e.g. a teacher) would mean to determine the merit of the products of its

functioning (e.g. students' achievements). In recent years some interesting attempts have been made to extend the scope of the variables of evaluation in various educational models (Stake, 1967; Stufflebeam, 1969, 1974; Stufflebeam *et al.*, 1971; Alkin, 1969). Stufflebeam's CIPP Model, possibly the most elaborate evaluation model (Stufflebeam, 1969; Stufflebeam *et al.*, 1971), suggests that evaluation focus on four variables of each evaluation object: its goals, its design, its process of implementation, and its outcomes. According to this approach, to evaluate an educational project, for example, would mean that one would assess the merit of its goals, determine the quality of its plans of action, assess the extent to which those plans are being implemented, and finally determine the worth of its outcomes.

5) What criteria should be used to judge the merit of an evaluation object?

To choose the criteria that should be used to judge the merits of an evaluation object or its variables is one of the most difficult tasks in educational evaluation. Those who perceive evaluation as a process intended to determine whether goals have been achieved (Tyler, 1960, Provus, 1971) make this task easy for themselves by ignoring the whole issue of evaluation criteria. Actually, they use "goal achievement" as an evaluation criterion to judge the merit of the object without justifying whether it is an appropriate criterion and without considering any alternative criteria. Alternative criteria might be responding to identified needs, ideals or social values, or comparison with other objects. The criterion (or criteria) to be used for the assessment of a specific object must be determined within the specific context of the object and the function of its evaluation. While in most cases the evaluator does not have the authority to choose among the various alternative criteria, it is his responsibility to ensure that such a choice be made and that he be able to provide a sound justification for such a choice.

6) Who should be served by evaluation?

If evaluation is to be useful to all, it has to be useful to some specific audience or audiences. While it is impossible to determine which is the most appropriate audience for evaluation, it is important that the specific audiences for each evaluation and their evaluation needs be identified in the process of planning the evaluation. The evaluation design should attempt to respond to those needs. It is also important to realize that various audiences might have different evaluation needs; thus, an evaluation

which responds to the needs of one audience might not respond to the evaluation needs of another audience.

7) What is the process of performing evaluation?

The process of evaluation may differ according to the perception guiding the evaluation. Some evaluations will follow a process of stating objectives, developing instruments, collecting data, interpreting findings, and making recommendations (Tyler, 1950). Others will use a process including identification of evaluation needs, collection of relevant information, and providing evaluative information to evaluation audiences (Stufflebeam, *et al*, 1971). While there seems to be no agreement among evaluators regarding the "best" process for conducting evaluation, most agree that evaluation should include a certain amount of interaction between evaluators and their audiences at the outset of the evaluation to identify evaluation needs, and at its conclusion to communicate their findings. Evaluation cannot be limited to the technical activities of data collection and analysis.

8) What are the methods of inquiry to be used in evaluation?

While challenging the usefulness of various research methods for evaluation studies (Stufflebeam, *et al*, 1971), recent years have seen a variety of new methods of inquiry introduced into the area of evaluation. In addition to traditional experimental and quasi-experimental designs (Campbell & Stanley, 1963), qualitative methods (Guba, 1978; Patton, 1980), jury trials (Kolf, 1975), system analysis, and many other methods have become legitimate methods for the conduct of evaluation. While some methodologists are still advocating the superiority of certain methods, there seems to be greater consensus among evaluators for the use of a more "pluralistic" approach regarding evaluation methodology.

9) Who should perform evaluation?

The type of evaluator is another consideration in conceptualizing evaluation. As they have crystallized into a professional group, evaluators have devoted a lot of attention to the characteristics of "good evaluators" and appropriate ways of training them. The discussion of such characteristics would be beyond the scope of this paper. However, in the context of developing an evaluation design, two kinds of distinctions should be made regarding the type of evaluators most appropriate for a specific evaluation. The first is the distinction between an internal evaluator and an external evaluator.

The other distinction is between a professional evaluator and an amateur evaluator. The internal evaluator of a project is usually one who is employed by the project and reports directly to its director. It is obvious that his objectivity and external credibility may be much lower than those of an external evaluator who is not employed directly by the project director and enjoys a higher degree of independence. An amateur evaluator (Scriven, 1967) is usually one whose major professional training is not in evaluation and whose involvement in evaluation represents only part of his job description. While his evaluation skills might be lower than those of a professional evaluator, he might have a good understanding of the project's unique evaluation needs and be able to develop good rapport with the members of the evaluated project. If an evaluation team is to be employed by a project, a combination of various types of evaluators should be encouraged.

10) By what standards should evaluation be judged?

Many attempts have been made during recent years to develop standards for the evaluation of educational activities. Boruch and Cordray (1980) analyzed six sets of standards and reached the conclusion that there was a large degree of overlap and similarity among them. The most elaborate set of standards and the one based on the broadest amount of consensus, is probably the set published by the Joint Committee on Standards for Educational Evaluation (Joint Committee, 1981). These standards were developed by a committee of seventeen members, chaired by Dr. Daniel Stufflebeam, which represented twelve professional organizations associated with educational evaluation. Thirty standards were suggested by the Joint Committee, divided into four major groups: utility standards, feasibility standards, propriety standards and accuracy standards.

SUGGESTED EVALUATION MODEL

The evaluation model of the project will be described below, in three sections. The first section will deal with the organizational structure of the evaluation mechanism, including role definitions of those involved, their responsibilities, authority and administrative subordination. The second section will describe the research design for the evaluation, its measurement tools, data collection methods and data analysis. The third section will describe a suggested system for communication between the evaluation team and the other bodies involved in the project.

The Organizational Structure of the Evaluational Mechanism

The evaluation mechanism of the project is composed of two major bodies: the internal evaluation team and the external evaluation team. Those involved in this mechanism and their integration into the structure of the project are shown in Figure 1.

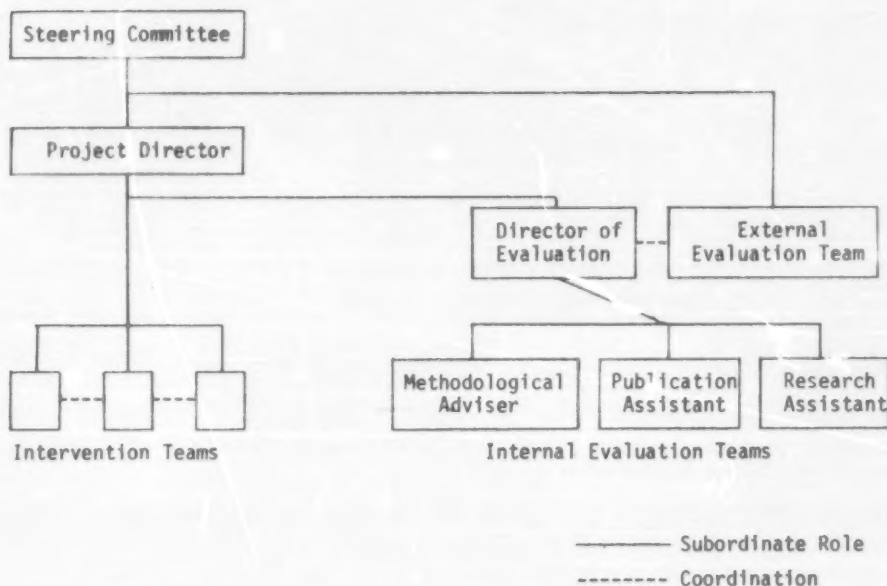


Figure 1. The Organizational Structure of the Evaluational Mechanism.

The internal evaluation team is the central body involved in the evaluation mechanism of the project. It has the major responsibility for the formative function of the evaluation of the project and secondary responsibility for the summative function of the evaluation of the project. The team is headed by the evaluation coordinator, who is appointed by the director of the project, and to whom the evaluation coordinator is directly subordinate. The evaluation coordinator is also responsible for coordinating the work with the external evaluation team. The internal team includes the following personnel:

- A) A research assistant, whose function is to help the evaluation coordinator in the gathering of data in the field and in the statistical processing of the data. B) A publication assistant whose function it is to gather data from

existing documents, to prepare bibliographic reviews based on relevant publications, and to assist the evaluation coordinator in preparing evaluation reports. C) A methodological adviser whose function is to aid the evaluation coordinator in the development of measurement instruments, the formation of research designs and the processing of data. The internal evaluation team is described here in terms of functions and not in terms of the number of full-time or part-time personnel needed for carrying out these functions. The number of positions needed will of course be determined by the size of the project and the resources available to it. For a project similar in size to the Helping System Project (HSP) in 1976, one can estimate the personnel needs as follows: the evaluation coordinator, a half-time position; the research assistant and the publication assistant, each between a one-quarter to a one-third position; the methodological adviser, between three to five hours weekly.

The external evaluation team should be composed of two or three members, at least one of whom should be an expert in evaluation and the others expert in the types of problems with which the project deals. The external evaluation team should be appointed by the steering committee (or by a similar body with authority to supervise the project and to determine its policy), in consultation with the project director and the evaluation coordinator. The team should be directly subordinate to the steering committee, from which it should derive its authority and to whom it should report its findings. The external evaluation team is the one primarily responsible to fill the summative function of the evaluation, but is also likely to fill a certain role in the formative function, as will be outlined below. The contract with the external evaluation team should be for a specified period, with one year likely to be the optimal time for such a contract. The sequence of actions which are to be taken by the external evaluation team may be described in terms of the following stages:

A. Orientation. In this stage, the external evaluation team meets with the steering committee and the project team. It studies the various documents relating to the project and observes the project in action. This stage is terminated by the submission of an outline of the team's action program for the year to the steering committee, which should lead to the signing of an agreement between the two. This agreement should detail the tasks of the team, its authority and the resources available to it.

- B. Organizing Data Collection.* In this stage, there are meetings between the internal and external evaluation teams, with the aim of determining the type of information the external team will need and by when the internal team should supply the data. Thus the internal team will be the body gathering the data for the external team. The latter is limited in its ability to perform such a task, and is restricted primarily to impressionistic reactions to what it sees in the field. A considerable part of the information the internal team gathers for itself can, of course, be used by the external team as well, but the internal team is likely to be asked to gather additional data which only the external team will need.
- C. Analysis of the Data and Observation of the Project.* In this stage, which continues throughout most of the year, the external team meets a number of times in order to analyze the data supplied to it by the internal evaluation team, in accordance with the agreed-upon timetable. The external team also observes the project in action a number of times, gathers impressionistic information as and if needed, and asks the internal team to gather additional information, beyond whatever it had originally requested. At this stage there might be meetings between the external team and the project team or the external team and the project head, during which the external team may offer feedback as formative evaluation to the project team.² During this stage the external team also meets with the steering committee to submit an interim report, either orally or in writing.
- D. Preparation of Summative Evaluation Report.* Toward the end of the year, the external evaluation team meets in order to analyze the data which was made available to it and/or the data it itself gathered during the year, and to summarize its findings in writing. The final report of the team should also include an evaluation of the internal evaluation team, where the external team serves in the role of the meta-evaluator³ of the internal team.
- E. Presentation of the Concluding Report.* A draft of the concluding report is submitted to the project head and to his evaluation coordinator, and both are given the opportunity to comment on the content and recommendations of the report, but the external evaluation team reserves to itself the final editing of that report. The final text of the report is submitted to the steering committee, and this submission includes a discussion about the recommendations and about the possibility of carrying them out

in practice. If there is need, the project head is permitted to submit an appendix together with the external evaluation team report, in which he details his reaction to the report. If the project team head does indeed submit such an appendix, and if the steering committee and/or external evaluation team decide to publicize the external evaluation report, the project head's comments should appear as an appendix to the external evaluation team's report.

The resources needed by the external evaluation team depend, of course, on the size of the project to be evaluated. In the case of the HSP, one may estimate that in 1976 the project needed the services of two or three experts who were employed as advisers under a special contract for fifteen to twenty days of work during the course of the year.

To summarize, the organizational structure suggested here attempts to deal simultaneously with the formative function and the summative function of the evaluation of a project, while ensuring the judicious use of the internal evaluation team and increasing the reliability of the summative evaluation by employing an external evaluation team. One may surmise that any project which utilizes this pattern of evaluation will fill the psychological function of evaluation by offering a "personal example" of controlled activity and will use evaluation as a force to spur on the project team toward grappling directly with the problems it faces.

THE RESEARCH DESIGN OF THE EVALUATION MECHANISM

The evaluation pattern proposed here presents a conceptualization based primarily on the special character of a project whose evaluation requires a wide range of research methods and measuring instruments. A major difficulty in locating suitable research instruments in the proposed approach is the fact that it attempts to set up evaluation activities which are meant to produce information which will meet standards of scientific accuracy on the one hand, while at the same time being practical and relevant to those engaged in the project. Unfortunately, there is a great amount of contradiction between the two, and the desire to satisfy both forces one to find some type of compromise solution which will result in an optimal balance between them. Another problem with the proposed approach is that it attempts to deal with a very wide range of evaluation activities both in terms of the evaluational functions which it wishes to fill and the number of variables

with which it wishes to contend. If one wants to propose an evaluation scheme which is not limited to measuring only those variables which have "approved" measurement instruments available, one must also propose ways to go about gathering data for the other variables with which he wishes to deal. In the following section we will try to suggest a wide range of research instruments, with varying levels of reliability and validity, which, in the writer's opinion, can be the "tools" of the members of an evaluation team and which can suggest ways for the development of new tools. These tools will be presented in relation to the four variables with which evaluation must deal.

A. Goals of the Project. The goals of the project must be evaluated, regardless of whether one sees their fulfillment as being the central measure for the evaluation of the entire project (Tyler, 1950) or whether one wishes to evaluate the project independently of its stated goals.⁴ In both cases this evaluation should help the project team formulate its goals and evaluate their quality. Three major factors must be taken into account when evaluating the goals of an educational project:

- a) The educational and social needs of those involved in the project, both those who will be benefiting from the project and those who will be "tendering services".
- b) The social norms and values of the society within which the project is to function, which affect the setting of priorities as to the investment of resources in various educational activities.
- c) The financial, social, political and national security constraints which affect the setting of priorities at the time that the project is to function.

It is almost superfluous to note that in most instances there is no total agreement between the three factors, and one must seek the optimal balance in devising the goals of the project. It is not the duty of the evaluator to determine how this balance is to be achieved, but it is his duty to provide those making such decisions with the information necessary in order to ensure that the decisions taken be as rational as possible. In order to fulfill this function, the evaluation team has available to it a number of techniques, of which it should choose those which it thinks are most suitable for the specific project involved and the possibilities open to the internal evaluation team. A number of these techniques are listed below, and in each case mention is made of where further information may be obtained about that particular technique.

1. Review of Official Documents. The gathering and analyzing of various official documents (government decisions, laws, committee reports, resolutions by directors-general, the minutes of meetings, etc.) and of information sources (reports of the office of statistics, administrative reports, student files, etc.) can serve as vital information as to the needs and the system of norms and values by which the goals of the project should be evaluated. The techniques of "content analysis" (Berlson, 1954) and descriptive statistics can be important tools for the evaluator.
2. Review of Research Literature. The results of research studies which have dealt with topics relevant to the particular project or populations similar to the one involved (e.g. culturally deprived students) can supply important information for the formulating of goals for a project and for evaluating them. Systematized collections of educational research studies⁵ as well as basic journals in behavioral research can also be important sources for a systematic review of the literature.
3. Needs Assessment. Various techniques which were developed in recent years for gathering data regarding the order of priorities in the educational aims of those involved in various educational frameworks, such as parents, students, teachers and administrators. There are a number of instruments that have been developed for accomplishing this task, the most elaborate of which may be the one developed by the Center for the Study of Evaluation of the University of California in Los Angeles (Hoepfner and others, 1973).
4. The Delphi Technique. This technique was originally developed as a means for gaining consensus in a group about predicting future phenomena (Helmer, 1966), and has been used extensively in the past few years as an instrument for defining goals and determining standards for the evaluation of educational programs (Uhl, 1971; Weaver, 1972).
5. Field Surveys. One of the accepted research methods in the behavioral sciences which the evaluator may use is the conducting of a survey in order to ascertain relevant information which is not available from other sources. In recent years there has been renewed use of anthropological approaches in the conducting of these surveys (Sindell, 1969; Patton, 1980).

B. Plans. After the project team has determined its goals, it must consider which methods to adopt in order to accomplish these goals. While it is true that a complete assessment of the chosen plans can be done only after they have been in effect for some time, so that their influence in

achieving the project's goals can be measured, the internal evaluation team must nevertheless - even before putting these plans into effect - supply the decision-makers within the project with the required information so that they can choose the plans that are most likely to succeed. The external evaluation team must also consider the proposed activities to be undertaken and must evaluate the project - not only in terms of its implementation and accomplishments, but also in terms of the quality of the strategies it chooses for itself. A number of important tools are available to the evaluation team if it wishes to find out information about the quality of the plans of action.

1. Review of the Research and Evaluation Reports. The evaluation team should supply the project team with summaries of research findings which relate to the effectiveness of the various plans which the project team is considering adopting for the project. Those sources quoted in the previous section can aid the team in this regard as well.
2. Inspection of the Program by Experts. Alternative strategies may be investigated by experts in relevant areas of specialization. It is desirable for this evaluation to be carried out in a systematic manner by means of questionnaires, evaluation scales and checklists.⁶
3. Advocate Teams. This is a method that was developed recently (Reinhard, 1972) for formulating and evaluating new plans of action by teams of experts who represent different approaches to the given problem. This method is especially applicable for the formative evaluation of the development of alternative activity programs.
4. Laboratory Experiments. In certain cases the internal evaluation team can conduct laboratory experiments on a limited scale to examine different aspects of alternative strategies before the project team begins using them on a larger scale.

C. The Process of Implementation. A project which has set sound goals for itself and which has developed suitable plans of action to achieve these goals, must be concerned about the quality of implementing the plans which were developed. If a program is not implemented properly, it is meaningless to ask to what extent it has accomplished the goals set for the project. It is important to identify any difficulties there might be in implementing the program, so that these can be overcome in time and so that the chances for success of the project can be increased. In order to evaluate the implementation of the project, the following methods are available to the evaluation team:

1. Administrative Control. Systematic weekly reports by the team members of their weekly activities, noting the hours of contact between the team members and the clients, the hours of contact among the team members, the calculation of the time devoted to implementing various activities, the classification of the products (like reports, working papers, etc.) produced by the team members - can all serve to supply useful information about the implementation process of the programs of the project. The organization and systematic analysis of this information by the evaluation team can be an important basis for the evaluation of the process. The use of the PERT system (Cook, 1966, 1971) or similar techniques can be most useful in this regard.
2. Observations. Structured observations (using observational tools) and unstructured observations (using supervisory and observational teams) of the activities in the field are an important source of evaluational information about the implementation of the project. The professional literature suggests numerous examples of observation tools and ways to analyze learning activities and interaction in small groups (Bales, 1950, Simon and Boyer, 1968), even though in many cases the evaluation team will have to develop its own unique observation tools in accordance with the needs of the project (Nevo, 1976).
3. Questionnaires. The distribution of attitude feedback questionnaires (Edwards, 1957) to the various parties involved in a project (students, teachers, parents, administrators) can serve as a source of information of the degree of satisfaction of the various parties involved in the project with its activity methods, and can point out shortcomings in the activity process.

D. Products. This group of variables relates to the influence the project has upon the target population both in terms of the achievement of the goals which the project set for itself and in other ways which are not directly related to the goals of the project. As was mentioned earlier, this is the only variable with which most evaluational designs concern themselves. The approach presented here feels that not only should evaluation not be restricted to these variables, but that on occasion one should not invest any resources at all in evaluating the impact of the project, if its goals, plans of action and implementation process do not show that the project is functioning properly. On the other hand, if the evaluation shows that the project is indeed functioning properly, one should not restrict himself to evaluating the variables mentioned above, and should make every effort to evaluate the project's accomplishments and its impact upon the target population.

The definition of the outcome variables is not easy, especially in such projects as the HSP, which aim to accomplish sweeping changes in many fields, and which do not restrict themselves to trying to influence only specific areas in regard to the target population (such as the achievements of the students, a change in the pedagogic methods and so on). The first source for defining the outcome variables can be the goals of the project which had already been defined earlier, but the evaluation should not be restricted solely to those variables which are directly connected to the declared goals of the project. The evaluator must also investigate other variables which he decides deserve to be examined, based on theoretical and empirical considerations. For example, in evaluating the achievements of students in a project for the teaching of mathematics, one should also evaluate the students' achievements in the affective domain, even if the project defined its goals only in the cognitive domain.

For each project one must define those outcome variables which are appropriate, and this is one of the important tasks of the evaluation team. Below we will bring an example of a number of such variables for the HSP. In Table 2 the variables are divided into three groups, and for each variable appropriate measurement instruments are indicated.⁷

The changes in the variables below as a result of the implementation of the project should be organized by means of a quasi-experimental design (Campbell and Stanley, 1963), such as the Non-Equivalent Group Design⁸ in its "weaker" design without random assignment of the intervention variable. In this research design one must use groups similar to those which are involved in the project as a control group. Both in the experimental group and in the control group, one must measure the outcome variables once before the project begins and once after a sufficient time period has elapsed to have allowed for the project to have had an impact on the community and its educational institutions. It may be that a period of two years, accompanied by evaluation of the implementation process of the project, would be the minimal amount of time needed in order to measure the influence of the system upon the community where the project is instituted.

It is clear that the suggested research design is not a strong design which can provide "incontrovertible proof" of the effectiveness of the project during the time that it is active in the community, but the use of a control group and the accompanying measurement of variables within the evaluation of the implementation process can provide important information about the changes which take place in the community and the connection between these changes and the activities of the project.

Table 2. An Example of Outcome Variables for Evaluating the Helping System Project.

Variable	Measurement Instruments
<u>1. The School</u>	
1.1 Educational Achievements	Standardized Achievement Tests in the Major Subject Areas
1.2 Drop-out Rate	Student Lists
1.3 Success in External Examinations	Matriculation Examinations, School-Leaving Examinations, etc.
1.4 Streams and Educational Levels	Lists of Students in Special Education Stream, Vocational Streams
1.5 Educational Background and Experience of Staff Members	Teachers' File or Questionnaire
1.6 Teacher Turnover	Teachers' File
1.7 Student Satisfaction	Attitude Questionnaires to Students
1.8 Teacher Satisfaction	Attitude Questionnaires to Teachers
1.9 The Quality of the Curriculum	Examination of the Curriculum in Actual Use
<u>2. The Local Community</u>	
2.1 Satisfaction of Parents with the Educational System	Attitude Questionnaires, Interviews
2.2 Satisfaction of Local Residents who are not Parents	Attitude Questionnaires, Interviews
2.3 Parent participation in Educational Funding	Examination of the Budget of the Educational Institutions
2.4 Parent Involvement in Educational Decision-Making	Observation of Parent Committee Meetings
2.5 Community Involvement in Educational Decisions	Analysis of the Minutes of Meetings of the Local Council
2.6 Level of Cultural Activity in the Community	Examination of the Participation in Cultural Events, Use of Libraries
<u>3. Local Leadership</u>	
3.1 Turnover of Personnel	Workers' File
3.2 Interaction between Personnel of Different Levels of the Local Educational System	Interviews with Various Personnel Members
3.3 Interaction between the Local Leadership and Ministry of Education Personnel	Interviews
3.4 The Quality of the Discussions of Meetings of the Local "Elite"	Observation of these Meetings
3.5 Special Initiatives of the Local Leadership	List of Special Projects and Events
3.6 The Use of Proper administrative Procedures	Examining Record Keeping Procedures, Files, Information Pools

COMMUNICATION BETWEEN THE EVALUATIVE MECHANISM AND THE OTHER BODIES INVOLVED IN THE PROJECT.

The existence of open lines of communication between the evaluative mechanism and the various parties involved in the project is an important condition for evaluation to be an effective tool in the operation of the project. Evaluative information which was not gathered by prior contact between the evaluation team members and the consumers will be self-defeating. If there is an accumulation of data which is not transmitted in a systematic manner to those who need it, it becomes a burden, rather than being a tool designed to aid the project. There are, of course, many ways to ensure communication. Below are listed a number of possibilities for the evaluation teams which composed the evaluation mechanism for the HSP.

The internal evaluation team must ensure that it has an intensive communication system with the project members. A number of possible communication avenues are available to the team within the framework of the project. These are listed below.

1. Personal meetings with individual members of the project team or with various groups within the team. These meetings must take place a number of times during the year. At the beginning of the year these meetings should be used to identify the evaluation needs of the members of the project team, while during the year they should be utilized for reporting on the data that has been collected and for identifying additional needs.
2. Participation in Staff Meetings. A representative of the internal evaluation team must participate in all the staff meetings as an observer, in order to identify evaluation needs as they surface during the deliberations, as well as to supply existing information which may help the staff during its deliberations. During the year it is advisable to devote a number of staff meetings to specific discussions about the evaluation data which will be presented by the internal evaluation team.
3. Information Bulletin and Bibliography. Once every so often, at fixed intervals (every month or two months) the evaluation team should prepare a short document for the project team, which should include a summary of the evaluational information which was gathered since the last bulletin was published. This bulletin should include empirical data and bibliographical surveys of interest to most members of the team as well as a description of evaluation activities which had been carried out in the recent

past and those which are planned for the near future. It is also worth considering circulating the bulletin among various groups which are connected with the project.

4. Specific Evaluation Reports. A number of such reports will be prepared during the year, and will be devoted to a comprehensive report gathered by the team regarding specific problems which had been defined in advance and to which the team had committed a good part of the resources available to it. This may include, for example, a report on social and economic problems of one of the communities of the project, the investigation of the interaction method accepted by the team, or a comprehensive bibliographical survey about a defined topic. It is advisable for such a report to be followed by a discussion among the staff.
5. Annual Evaluation Report. Once each year the evaluation team should prepare a summary report which will detail all its activities during the past year and will present the main points of its findings. This report should be based on the specific evaluation reports and on the information bulletins which were prepared throughout the year. This report should also serve the project director in the preparation of his annual report to the steering committee or to other parties not connected with the project.

The external evaluation team conducts most of its communication with the project through the steering committee, to whom it is subordinate in administrative terms, but it also maintains a great amount of direct communication with the project team. The communication means of the external evaluation team may be summarized as follows:

1. Ongoing Communication with the Project Director and with the Internal Evaluation Team Coordinator in order to coordinate the collection of data and to analyze its significance. In these contacts, the type of information that the external team wishes to obtain from the internal evaluation team will be determined, as will be the additional information which the external evaluation team wishes to gather on its own.
2. Meetings with the Project Team. It is advisable for the project team to meet every few months with the external evaluation team at which time the project team can receive feedback from an external body which is not involved in the day-to-day operation of the team. These meetings will permit the project team to utilize the external evaluation team for for-

mative evaluation, in addition to the summative evaluation which is entirely in the hands of the external evaluation team.

3. Meetings with the Steering Committee. The steering committee should invite the external evaluation team to a number of meetings in order to receive interim reports of the findings of the team before the annual report is submitted. These meetings will enable the steering committee to receive a preliminary report of the operation of the project and to examine to what extent the external evaluation team has gathered evaluational data which may be useful to the steering committee.
4. The Annual Evaluation Report. This report should be submitted to the steering committee as a written document to summarize the yearly operation of the project in a summative report. The first draft of the report should be submitted to the project director before it is submitted to the steering committee, but the authority for final editing of the annual report should be reserved for the external evaluation team. The possibility of having this report distributed among other parties not connected with the project and the steering committee should be agreed upon at the time the agreement is signed between the steering committee and the external evaluation team, when the work is yet to get underway.

SUMMARY

This article has attempted to suggest an evaluation model for the Helping System Project or similar projects of change and intervention within the framework of cooperation between university faculty members and local education communities. The suggested program is based on the analysis of the characteristics of the project and on the clarification of the significance of evaluation and its role within the framework of the educational enterprise. The evaluation program includes a system of principles and their translation into suggested activities within the plane of the organizational structure of the proposed evaluation mechanism, the research tools available to it, and its communication system.

In order to ascertain to what extent the suggested program is indeed suitable for evaluating the project, it must be subjected to the test of criticism and of empiric testing. To the extent that the project is an object for evaluation, its evaluation program must also be subject to evaluation. As a first stage, this program should be submitted to educators and evaluation experts who will evaluate the quality of its goals and the

- SINDELL, P.S. Anthropological Approaches to the Study of Education. *Review of Educational Research*, 1969, 39, 593-605.
- SMILANSKY, M. and SANDERS, D. *A Regional Helping System: A Way for Systematic Development of Educational Institutions to Meet the Needs of Individuals and Society*. Tel Aviv: School of Education of the Tel Aviv University, 1975 (in Hebrew).
- Sociological Abstracts*. New York: Sociological Abstracts, 1954 to date.
- STAKE, R.E. The Countenance of Educational Evaluation. *Teachers College Record*, 1967, 68, 523-540.
- STUFFLEBEAM, D.L. Evaluation as Enlightenment for Decision-Making. In: Walcott H. Beatty (Ed.) *Improving Educational Assessment and an Inventory for Measures of Affective Behavior*. Washington: Association for Supervision and Curriculum Development, NEA, 1969.
- STUFFLEBEAM, D.L., FOLEY, W.J., Gephart, W.J., GUBA, E.G., HAMMOND, R.L., MERRIMAN, H.O., and PROVUS, M.M. *Educational Evaluation and Decision-Making*. Itasca, Ill: Peacock, 1971.
- STUFFLEBEAM, D.L. The Relevance of the CIPP Evaluation Model for Educational Accountability. *Journal of Research and Development in Education*, 1971, 5(1).
- STUFFLEBEAM, D.L. Toward a Technology for Evaluating Evaluation. Paper presented at the 1974 Annual Meeting of the American Educational Research Association, Chicago, April 1974.
- TYLER, R.W. *Basic Principles of Curriculum and Instruction*. Chicago: University of Chicago Press, 1950.
- UHL, N.P. *Encouraging Convergence of Opinion, Through the Use of the Delphi Technique, In the Process of Identifying an Institution's Goals*. Princeton, N.J.: Educational Testing Service, 1971.
- WEABER, W.T. The Delphi Forecasting Method. *Phi Delta Kappan*, 1971, 52, 267-272.
- WEBB, E.J., CAMPBELL, D.T., SCHWARTZ, R.D. and SECHROST, L. *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally, 1966.

LATENT VARIABLES IN CAUSAL MODELS:

AN EXAMPLE FROM EDUCATIONAL RESEARCH

Itai Zak

This chapter deals with a new methodological development in data analysis: the linear structural relation model. In this model the parameters are estimated under least squares criteria. The basic notion of this model is explained and an empirical example is given in some detail, in which the model is applied to data concerning the retention of teachers in the teaching profession.

It is suggested that this newly developed methodological model may be found to be a powerful tool for widespread educational research, as well as to help solve evaluation problems in natural settings where a large pool of data is available, but the certainty about the statistical and conceptual behavior of the variables is not fully established.

Causal models with directly observed (manifest) variables, usually appearing under the heading of simultaneous equations, have been a principle tool in econometrics since the 1930s. Around the mid 1960s sociologists started to use path models in their research as well, borrowing the term from genetics (see Wright, 1934).

There has been some development of models with indirectly observed (latent) variables as well. Conceptually, the transition from manifest to latent variables in path modelling relied heavily upon classical models of factor analysis which have been developed and used by psychologists since the early 1930s. However, statistical methods that combine path models with latent variables were not developed until the early 1970s (Joresk y's LISREL versions of maximum likelihood approach, 1970-1978; Wold's Partial Least Squares (PLS) estimation, 1966-1979).

The PLS approach to path models with latent variables can be referred to

as "soft modelling" occupying a position between the "hard" statistical assumptions of traditional model building and data oriented approaches. As Wold (1979) has stated: "The PLS approach to path models with latent variables is primarily intended for multidisciplinary and other applications where the problems explored are complex and theoretical knowledge is scarce" (p.5).

The causal model which Duncan (1966) constructed on the basis of Turner's study of the determinants of aspiration can serve as a prototype of a linear, additive, asymmetric (recursive) causal model. As a convention, one-way, straight-line arrows connect independent (exogenous) and dependent (endogenous) variables. Unanalyzed relationships are depicted by two-headed arrows, with connecting lines which identify the relationships excluded from the model. The numerical quantities are values of path-coefficients (i.e. standardized regression coefficients).

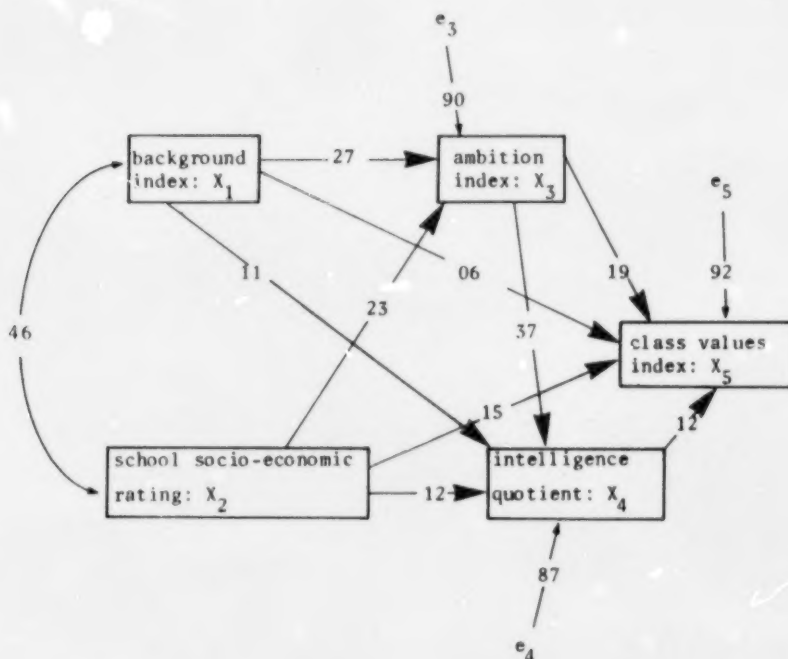


Figure 1: Illustrative path analysis (From Duncan, 1966)

The above illustrative causal model is a very basic example, where there are no unmeasured variables (except the residuals), the residuals are assumed to be uncorrelated and each of the dependent variables is directly connected to all preceding variables in the postulated causal sequence.

Especially pertinent from the perspective of causal modelling in educational sciences, is the extension of the path model to include unobserved latent variables. Whether the focus of analysis is individual behaviors within or across educational systems, or analysis at the systemic level, most of the relevant phenomena dealt with in education are complex, multivariate and not directly observable. Fundamental concepts of educational research such as child rearing, schooling, learning achievements, socio-cultural status, class environment, school climate, intelligence and the like are not directly observable. This suggests that many of the central independent, intervening, and dependent variables in educational analysis are fairly complex constructs which subsume and consist of various discrete or manifest indicators.

Only a few of these concepts are relatively established. One of the already established constructs is intelligence; and its indicators, the mental abilities, are referred to in Wold's (1979) terminology as reflective indicators. In other concepts, class environment, for example, the indicators serve as a measurement device for the complex phenomenon. These indicators are grouped into a block to form a construct, thus they are called formative indicators. As we shall see later, different models of relations between the latent variables and their indicators stem from the above definitions.

Latent variables in causal models become even more viable for educational scientists when it is recognized that many areas of inquiry explicitly cluster causal factors into general domains. For many independent variables (such as school achievements, educational and occupational attainments) the researcher would like to make generalizations about aggregate sociological, psychological and educational determinants, as well as to specify influences within the general domains. The combination of both manifest and latent variables in the same analysis lends itself to such questions. Crude verbal theories are compelled to be transformed into more exactly specified and explicitly articulated path models. Only the latter can be tested, refined and subjected to further empirical scrutiny and, even more importantly, can also be subjected to conceptual and theoretical evaluation.

In the field of education, especially in educational evaluation, true experimentation with random assignment of individuals to groups and groups to treatments, is very rare. Moreover, even when a manipulation of the independent variable is the practice, e.g. in curriculum evaluation, there is no real control of exogenous variables. In addition, the treatment itself is composed of several mutually related variables. Small samples (aggregates of students into classrooms or teachers into schools) with rich data are also a common practice, and the relevant variables are, at most, incomplete. Another problem is the low reliability of manifest variables. Manifest variables are often comprised of very few items on a questionnaire; sometimes even only a single question. Under such circumstances, there is a need for a new approach which takes into account the above difficulties. One of the methods which takes into consideration the problematic situation of a natural field study is the PLS method.

PLS PATH MODELS WITH LATENT VARIABLES - AN EXPOSITION

The conceptual design of a PLS model is provided by its arrow scheme and by its block structure. Figure 2 shows a graphical representation of a linear structural relation model (Noonan and Wold, 1977) and its results. We shall use this example to illustrate the various features of the model.

The model contains three types of variables:

- 1) The structural variables (latent variables) X , Y , Z are represented by encircled names (for convenience we use Latin, instead of Greek letters, as should have been done. Thus, instead of using the symbol $\eta = (\eta_1, \eta_2, \eta_3)$, for all structural variables, we use X , Y , and Z).
- 2) The indicator variables x 's, y 's and z 's are represented by names in rectangles.
- 3) The residual variables d and e , are represented simply by their notations.

Two parts of the model can be distinguished: the inner part, or the inner relations (the structural relations), and the outer part or the measurement models. The inner relations describe the relationships between the latent variables, and the measurement models describe the relationships between the latent variables and their indicators.

The inner part of the model in Figure 2 consists of the structural variables "home", "school", and "achievement". These are constructs which are not observed directly. The relations between these factors (another name for the latent, construct or structural variables) are given by path-

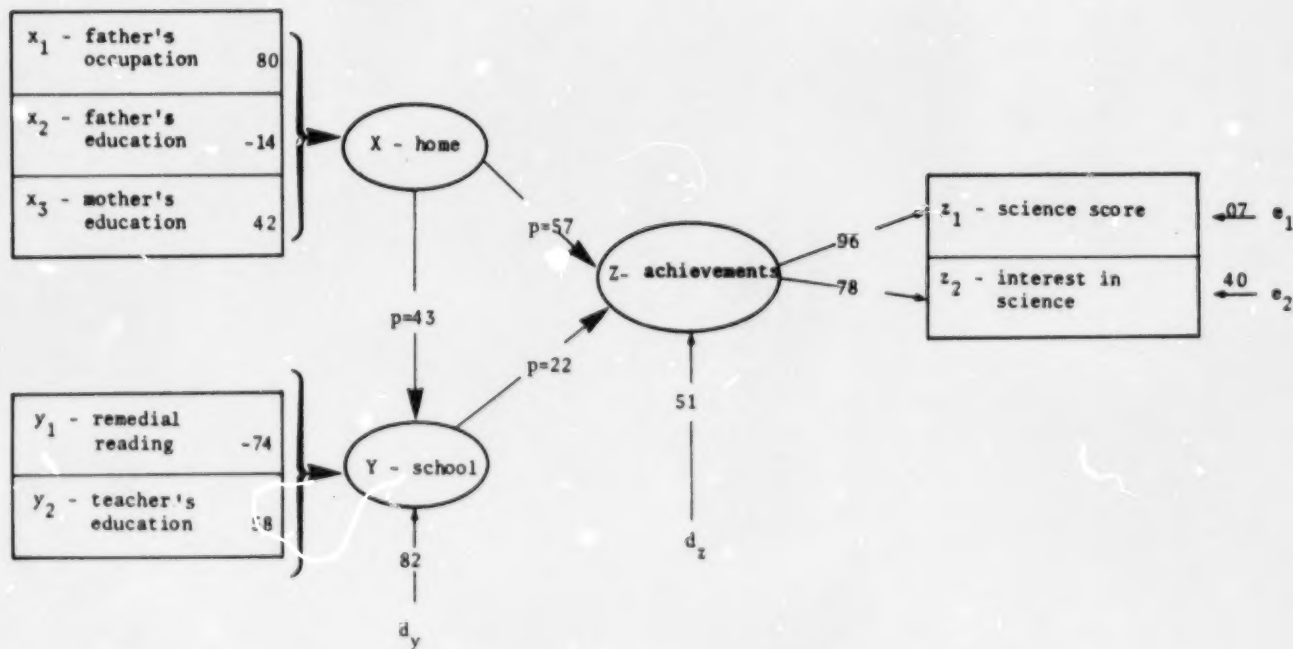


Figure 2: An example of a basic PLS model with latent variables
(adapted from Noonan & Wold, 1977)

coefficients (denoted by p 's). The relationships may be interpreted by sentences like: "Z is explained by X", or "Z is predictable from X", or "influenced", or even "caused", each according to the research design and its theoretical rationale.

The outer part of the model relates the factors X, Y, Z to their indicator variables. The indicator variables themselves are partitioned into blocks and each factor is measured by variables of only one block. (An extension of the model, in which several factors can be measured by the same block of indicators can be found in Lohmoller (1979). An illustration of such results will be given later in this chapter.) What is of importance are the two types (or modes) of outer models, the inwards (I) and the outwards (O) directed, measurement model.

In the outwards directed block, it is assumed that the factor is a predictor of its indicators (such as in Factor Analysis). In this case, the observables are assumed by the researcher to be quantified aspects of a unifying concept, thus the indicators are reflective. This is illustrated by the latent variable Z in the model. Subsequently, the factor (the latent variable) accounts for only part of the variance of its indicators (the common factor variance), and there is a measurement error; a residual variable d_j .

In the inwards directed blocks, the factor (X and Y in our example) is regressed on its indicators. That is, the latent variable is a transformed composite variable of a set of indicator-variables, and there is no residual variable, or measurement error in this measurement model. All the variance existing in the set of the indicators is transformed to the composite (latent) variable.

The coefficients in the outward model (Z) are, in a wider sense, factor loading (or correlations of the indicators with the factor) while in the inwards directed model (X, Y) they are regression coefficients.

There are differences between the inwards and outwards directed measurement models with regards to all estimated parameters. For instance, the correlation between the structural variables is highest under the situation in which both measurement models are inwards directed (such as between X and Y). Such a model resembles canonical analysis in which the correlation between the two sets of variables (actually, between two composite scores) is maximized. In our example, this correlation is .43.

When we are not interested in the most predictable criterion, but would like to predict our real criterion variables, the dependent latent variable is outward directed (such as in Z). In such a case, the correlation between the two latent variables (X and Z) is lowered, but the variance explained (the average sum of squares of factors loading of both blocks) is higher. This model resembles the canonical redundancy model (Van der Wollenberg, 1977).

The highest explained variance (and the lowest correlation between the latent variables) will appear in using the outwards directed model for both latent variables. This model resembles the interbattery factor model (Tucker, 1958), and is the convention in Joreskog's LISREL model. (For details, see Lohmoller, 1979).

Thus changing the direction of a measurement model can change the estimates of the parameters. Wold's PLS permits such changes, which make the model more flexible, and highly related to the real data. When one is uncertain of his latent variable (lv) as an already established construct underlying its indicator-variables (and when the lv is exogenous), one uses the inwards directed measurement model. When the construct has such properties of an established construct (and/or is endogenous), the outward directed model is considered, as it is depicted in Figure 2 for pupils' achievements.

Several generalizations about the implementation of PLS should be noted:

- 1) PLS can be used both for time series and cross-sectional data. The arrow scheme is the postulated hypothesis, as derived from the theoretical rationale.
- 2) The structural relations are derived directly from the arrow scheme according to the chosen measurement model and on the basis of (a) maximizing the correlation between latent variables that are directly connected by an inner relation (specified by the researcher), b) minimizing the residual variances of the inner relations (i.e. endogenous variables), and c) minimizing the residual variances for all outwards directed blocks (i.e. maximizing the common variance of each block).
- 3) PLS estimation is an iterative procedure. The structural relations are estimated by ordinary least squares, with each regression generating estimates for a subset of the unknowns to be estimated. That is, one part of the parameters is considered as known and held fixed, while the other is estimated. (The term "partial least squares" stems from this partitioning of the parameters).

- 4) The interpretation of latent variables and relations is the same as for ordinary regression results. For example, the path from X to Z in Figure 2, which yielded a coefficient of .57, indicates the expected increase in student achievement levels if the background variable increases by 1 unit, while the school conditions variable remains constant, the indirect influence in path analysis terms of, say, "home" on "achievements" through "school" would be only .095 (i.e. $(.43)(.22)$), as it is usually interpreted by path analysis users.

The PLS approach has been applied in an impressive number of contexts, most of them interdisciplinary (for a list, see Hopple, 1979). The formal statistical model can be studied in Wold (1979).

In order to test the PLS model, it must be considered with empirical data. The following section introduces such an example.

TEACHER RETENTION IN THE TEACHING PROFESSION - A STRUCTURAL RELATION MODEL

The data set for testing teachers' mobility is taken from Horowitz & Zak, 1979. The data is comprised of a 12% stratified sample of elementary and intermediate (compulsory) schools and their teachers at work, in three educational districts in Israel. This data was collected in 1975 and followed by a survey of teachers' mobility in the same school units in 1976. In this sample, full data was secured on 61 school units.

The data to be analyzed in the present example is comprised of:

- 1) organizational characteristics of the schools, including personnel attributes; 2) teachers' educational ideology; 3) the organizational climate of the school as perceived by the teachers; and 4) various measures of retention of the teaching personnel, including affective behavior (job-satisfaction), cognitive behavior (overt tendencies to stay in the profession), and instrumental behavior (three measures of the teacher turnover rate)

The methodological problems in analyzing such a complex phenomenon are immense. There are many indicator variables (even after reduction procedures, 25 variables remained), and the sample is relatively small (61 schools).

In previous analyses (Horowitz & Zak, 1979; Zak & Horowitz, 1980; Zak 1981), the data was subjected to different statistical treatments in which some of the variables were related to some others, in canonical correlation analyses, or by cluster analysis techniques in which schools were grouped into clusters and several profiles of schools were detected. Yet, attempts to generalize

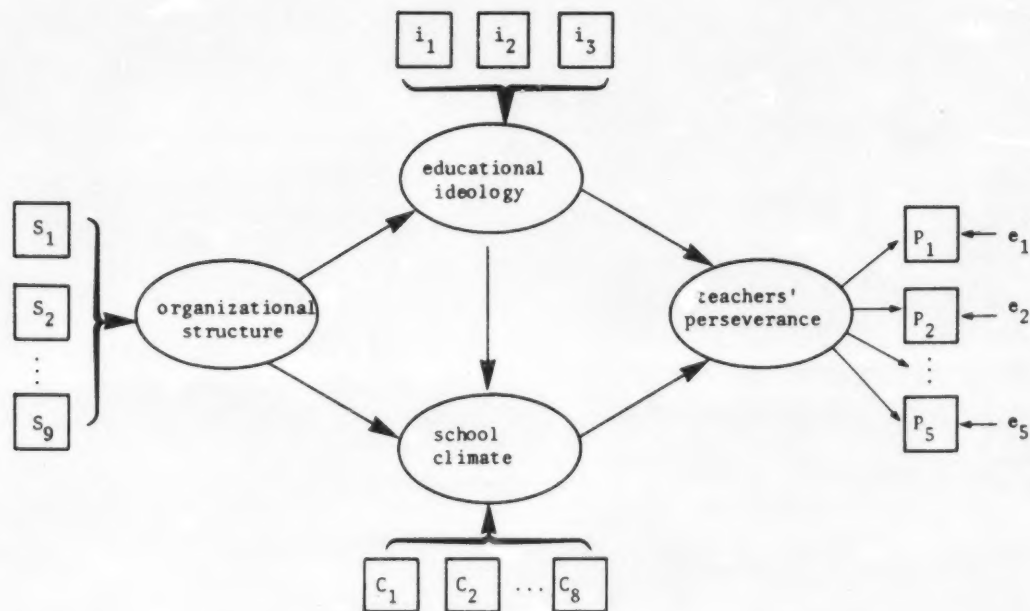


Figure 3: The schematic four-block PLS model for teachers' perseverance - Model I

the results into a systematic order encountered methodological and statistical difficulties. At most, a verbal generalization was possible, in which the several hypothetical constructs were related to each other.

The focal challenge, then, is to achieve enough understanding of the structural relations between the various constructs, without losing sight of the observed variables. That is, the aim of subjecting the data to a PLS model was to achieve some conceptual generalization while keeping our feet on the ground.

Figure 3 assumes the conceptual definition of the perseverance model of teachers on schools. We start with a simplistic schematic presentation of the model.

The model in Figure 3 is a casual one, which hypothesizes that predetermined factors are usually given for an individual school do not influence teachers' perseverance directly. The school's educational ideology and its climate, both as perceived by its teachers, serve as intervening constructs between the predetermined and the outcome constructs. That is, the notion that the organizational structure of schools influences teachers' perseverance directly is checked by an alternative model which assumes perceptual process constructs. The reason for challenging the "direct influence" notion rests on educational practice more than on rigorous theoretical background. There have been many attempts to change the environments of schools: to introduce educational innovations, to elevate teachers' autonomy, to increase peer relations by team teaching or to equalize teaching load. All the aforementioned concepts are perceived here as dimensions of the school's climate. School administrators make the effort to change school environments with the belief that climate makes a difference.

Attitudes towards education are also conceived as influencing teachers' behaviors. It is believed that such behaviors do not only include teaching strategies, but also include satisfaction with the job, tendencies to leave the teaching occupation and actual dropout from the job. In other words, socio-psychological processes within the school as a work place influence teachers' behaviors and may even override the effect the structural components have on them. Thus, the hypothesis that "subjective" constructs, such as the perceived ideology and climate, are mediating constructs between the input and the output, has practical implications as well as theoretical value.

The first model tested was a fully recursive one in which the path leading from S to P was also introduced. Elimination of this path was possible since

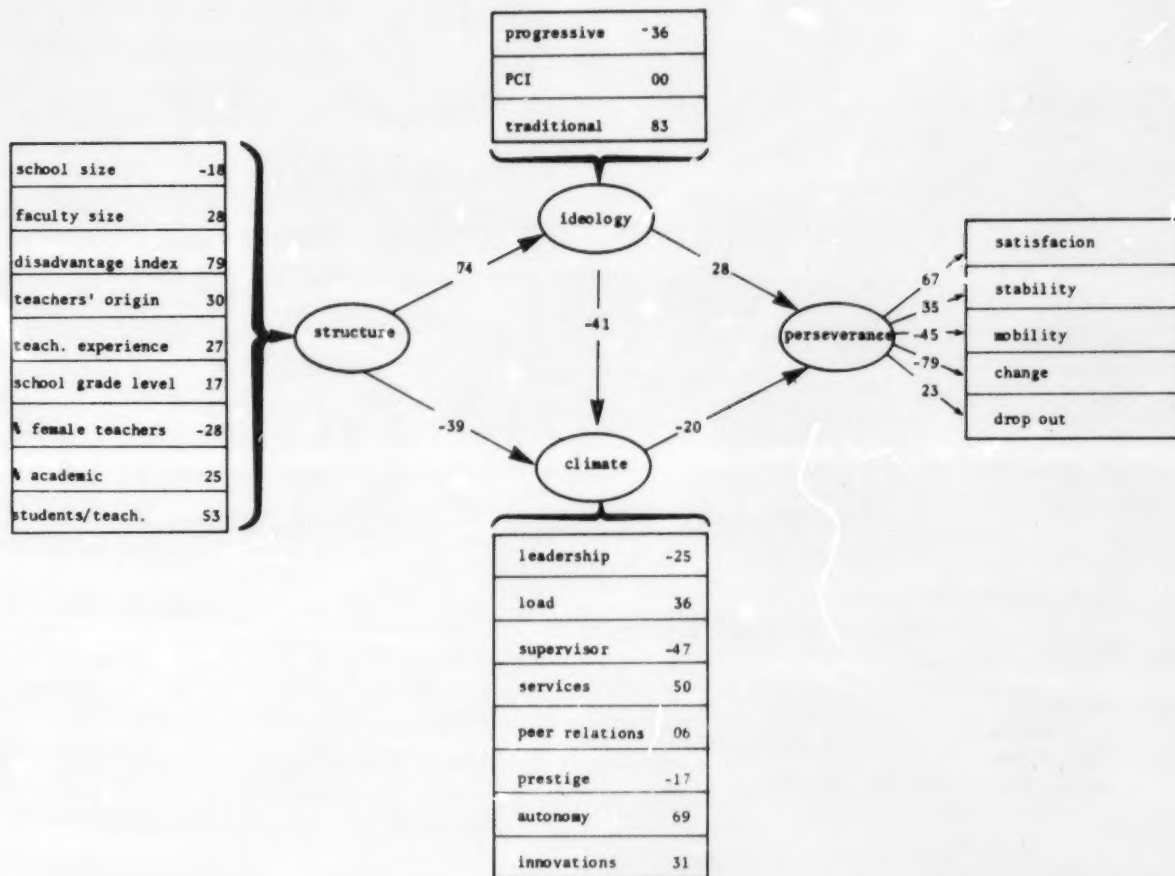


Figure 4: A reduced one dimensional model I of teachers' perseverance

its coefficient was negligible ($p = .01$ only). Only after the refutation of the alternative hypothesis (that S influences P directly), could model I then be tested as presented in Figure 3.

Model I

The results of Model I (as resulted from a new computation in its reduced form) are given in Figure 4 and in Table 1.

Table 1: Correlations and Structural Relations (Path Coefficients), Squared Multiple Correlations and the Explained Variance (MS Factor Loadings) of the Perseverance Model I ($N = 61$).

Structural Variables	Correlations (x 100)				Path Coefficients			R ²
	S	I	C	P	S	I	C	
S - organizational structure	100	-	-	-	-	-	-	
I - educational ideology	74	100	-	-	74	-	-	54
C - organizational climate	-69	-69	100	-	-39	-41	-	55
P - teachers' perseverance	35	42	-39	100	-	28	-20	20
Explained variance (%)	13.4	52.1	13.8	29.3				

The latent variables S (structure), I (ideology), C (climate) and P (perseverance) are encircled. Latent variable S is exogenous and fully determined by a weighted sum of its respective indicators; that is, S summarized the organizational structure as derived from the nine observables of its block (in the rectangle). The measurement model for the latent-variable (lv) S is then inwards directed, and the coefficients attached to each observed indicator are, in fact, standardized regression coefficients, or relative weights of the indicators for the composite score of S.

The intermediate constructs of Ideology and Climate are also measured by inwards directed models. This measurement model was preferred since Climate and Ideology are not well established constructs, one of the requirements needed for an outwards directed measurement model.

The perseverance lv is outwards directed as required and the arrows from the lv to the indicators symbolize this. The relative weights of the indicators on their respective lv's can be seen in the figure. The organizational structure is composed mainly of the disadvantaged index of the school and of student-to-teacher ratio, while the other indicators have less impressive

weights on it. The ideology is influenced mainly by the traditional attitudes toward education which the teachers held. The climate is composed of the dimensions of autonomy, services and the supervisor's role (negatively), and to a lesser degree by the dimensions of load and innovation adoption.

The perseverance lv is defined by its factor loadings (i.e. outwards directed measurement model). The highest loadings are the percent of teachers who changed schools the year after the survey, teachers' satisfaction with their job and the percent of teachers who were mobile (upward mobility in the school system, but not in their own school). Thus far we have dealt with the composition of the various lv's and the relative weights of their indicators.

The model can now be checked for its structural relations (i.e. the inner relations between the lv's). The impact of the exogenous lv S on I is very high (path coefficient of .74), while on C it is moderate ($P = -.39$). That is, a highly traditional ideology is developed (or exists) within schools which are characterized by the organizational structure as defined earlier.

The same structure affects climate. The negative sign attached to the path coefficient between S and C implies a school structure which is characterized mainly by the disadvantage index, generates a climate which is perceived as less autonomous, with less services, but with higher regard for the supervisor's role. This pattern of climate is enhanced by the traditional ideology of the school ($p = -.41$) as well.

According to the results in Figure 4, Perseverance is influenced directly by Ideology and Climate (p 's of .28 and -.20 respectively), and only through them, by the Structure. As recalled, the direct effect of the organizational structure on the perseverance lv was only .01, a very substantial reduction from the simple correlation between them (which is .35). The highest indirect effect of S on P is through I. Of the total effect of the organizational structure on the perseverance lv, 60% is transmitted via the school educational ideology, and the rest is transmitted through climate (22%) and via the more complex path of S through I to C on P (18%).

Yet, does the model fit the data? One possible examination of this question is the variance accounted for in the dependent variable. The squared multiple correlation of P on C, I and S is .196. That is, about 20% of the variance in perseverance is accounted for by its preceding lv's in the model. (No formal statistical test of significance is offered, due to the lack of the knowledge about the distributional properties of the variables).

Another possible check is the common factor variance of each lv (i.e. the explained variance of the outer model). Besides the Ideology lv in which its indicators account for 52% of the common (latent variable) variance (see Table 1), the others are less satisfactory. The explained variance in S is 13.4%, in C it is 13.8% and in P it is 29.3%. These figures are evidenced also in the exposition of the indicators in each of the blocks. Many indicators seem not to "participate" in the formation of the various lv's. Can we not account for more of the variations in perseverance? Do we understand enough about the structure and climate of the schools? A check for a second dimension (such as a second canonical variate) reveals that much more could be accomplished if some of the lv's were divided into more than one factor each. Thus, we proceeded to a second model.

Model II

The second model assumed the same four blocks of observables in the same recursive order. However, the analysis allowed for two possible factors for each of the blocks: Structure, Climate and Perseverance (the Ideology block remained intact). That is, the indicators of the organizational structure were allowed to form two composites instead of one. (For details, see Lohmöller, 1979). Each indicator can contribute to both factors in a given block, though the factors remain independent (i.e. orthogonal). Figure 5 presents the four-block, seven-factor PLS model of perseverance in teaching in this manner. The total variance accounted for by the outer model was 36.7% (compared to 21.3% in the previous analysis), a much more satisfactory fit of the model to the observed data.

The full results of the models which are presented in Tables 2 and 3, and the major results in Figure 5, should now be self-explanatory. The results demonstrate the centrality of both climate factors in accounting for teachers' perseverance.

Instead of the diffused construct of climate which appeared in model I, one can detect two more clearly established dimensions of the school climate. According to its indicators the first one can be labelled as open vs. closed climate, and the second one as a supportive climate. The supportive climate affects teachers' tendencies to remain at their school and their satisfaction with the job ($p = -.43$), which is comprised mainly of teachers' overt tendencies to stay in teaching. Fewer teachers in a supportive climate left their schools either to seek a higher position or to seek other jobs altogether. The open vs. closed climate affects only the turnover rate ($p = -.29$). The

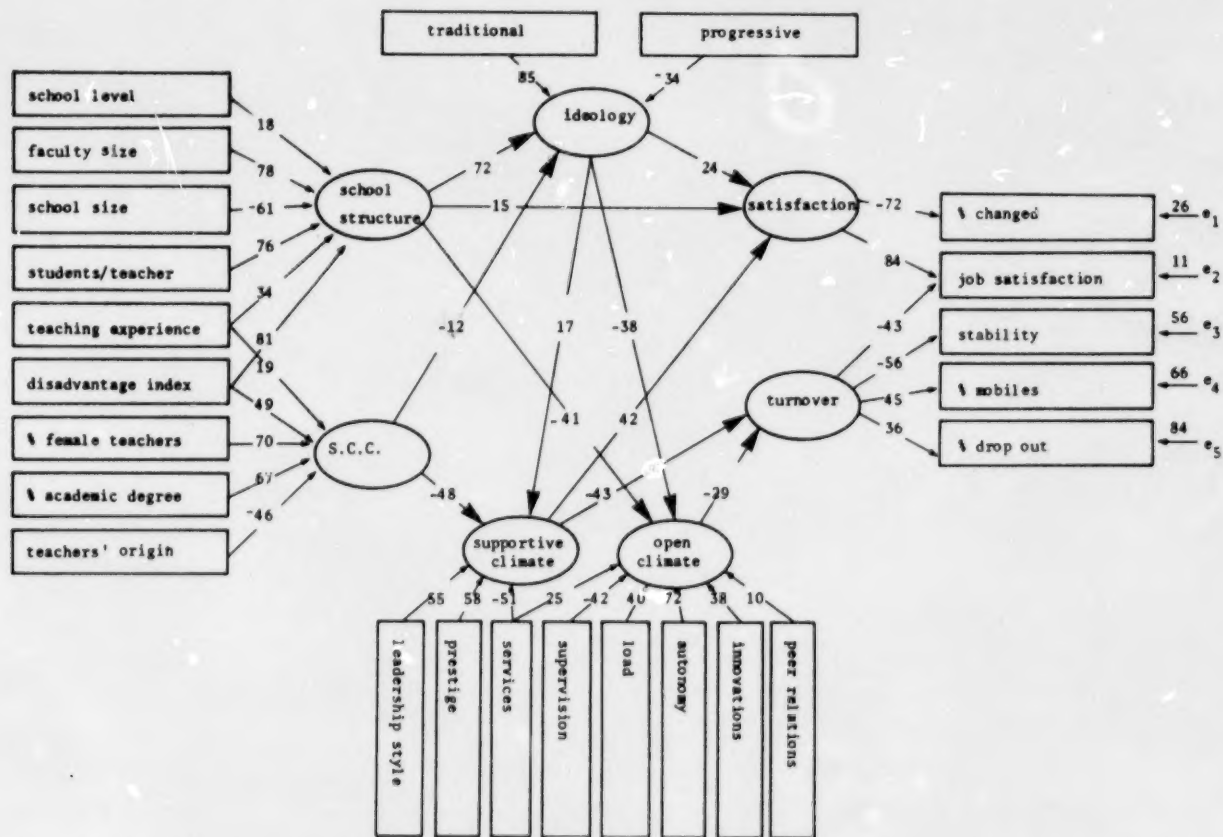


Figure 5: A four-block, seven factor PLS model II: schools organizational structure, educational ideology, school climate and perseverance of teachers in the teaching occupation (N = 61)
(Path coefficients ≤ 0.10 have been omitted. (See Table 2)

Table 2: Factor Correlations, Path Coefficients, Squared Multiple Correlations (R^2) of the Perseverance Model II and the Explained Variance (MS of Factor Loadings) (N = 61)

Structural variables	Correlations (x 100)							Path coefficients (x 100)					
	S ₁	S ₂	I	C ₁	C ₂	P ₁	P ₂	S ₁	S ₂	I	C ₁	C ₂	R ²
S ₁ - organizational structure	100							-	-	-	-	-	
S ₂ - socio-cultural status	0	100						-	-	-	-	-	
I - educational ideology	72	-12	100					72	-12	-	-	-	53
C ₁ - open vs. closed climate	-69	12	-69	100				-41	8	-38	-	-	56
C ₂ - supportive climate	-12	46	11	(05)	100			0	-48	17	-	-	24
P ₁ - satisfaction with teaching	35	-17	-37	-23	47	100		15	01	24	-04	42	33
P ₂ - turnover rate	21	29	17	-32	45	(-13)	100	9	-07	-07	-29	-43	32
Explained variance (%):	22.5	52.1		38.5	50.2								

Note: The correlation coefficients in parentheses are between the residuals of the structures, and not between the structures themselves, which are assumed orthogonal.

Table 3: Factor Weights and Factor Structure of the Perseverance Model II

Variable	weights (x 100)						
	S ₁	S ₂	I	C ₁	C ₂	P ₁	P ₂
1. No. of students in school	-61	-34	0	0	0	0	0
2. No. of teachers in school	78	05	0	0	0	0	0
3. Disadvantage index (students)	81	49	0	0	0	0	0
4. Teachers' origin (%)	18	-46	0	0	0	0	0
5. Teaching experience (years)	34	19	0	0	0	0	0
6. School grade level (1-6, 7-9)	18	04	0	0	0	0	0
7. % of female teachers	-04	70	0	0	0	0	0
8. Academic degree (% teachers)	15	67	0	0	0	0	0
9. Student-per-teacher ratio	15	14	0	0	0	0	0
10. Traditional attitudes	0	0	85	0	0	0	0
11. Progressive attitudes	0	0	-34	0	0	0	0
12. Pupil control ideology	0	0	-01	0	0	0	0
13. Leadership style	0	0	0	-07	55	0	0
14. Teaching load	0	0	0	40	-10	0	0
15. Supervisor role	0	0	0	-42	25	0	0
16. Services	0	0	0	34	-51	0	0
17. Teachers' relations	0	0	0	10	-02	0	0
18. Teachers' prestige	0	0	0	-03	58	0	0
19. Autonomy	0	0	0	72	-11	0	0
20. Innovation adoption	0	0	0	38	09	0	0
21. Job satisfaction	0	0	0	0	0	57	-62
22. Stability tendencies	0	0	0	0	0	23	-30
23. % of teachers' mobility	0	0	0	0	0	10	58
24. % of teachers change schools	0	0	0	0	0	-59	-41
25. % of teachers dropout	0	0	0	0	0	-10	-48

Variable	Correlations (x 100)						
	S ₁	S ₂	I	C ₁	C ₂	P ₁	P ₂
1.	22	-27	-08	-10	07	23	-04
2.	34	06	-02	-20	08	07	11
3.	77	13	60	-59	01	14	26
4.	49	-38	44	-41	26	13	06
5.	22	07	17	-05	-10	25	13
6.	-12	30	-05	07	04	14	-19
7.	-06	27	-30	12	-16	02	08
8.	-07	61	04	-02	-27	-28	06
9.	15	39	-02	-06	10	28	-08
10.	70	-08	95	-65	08	31	20
11.	-38	18	-60	42	-13	-30	00
12.	48	-21	54	-40	03	02	01
13.	-11	34	-18	36	71	23	-49
14.	-09	-11	-01	10	-13	-11	-05
15.	22	15	09	-22	38	30	-06
16.	-26	06	-41	42	-14	-13	-01
17.	-19	17	-24	43	26	06	-13
18.	-12	44	-14	28	70	26	-34
19.	-51	20	-54	77	16	09	-26
20.	-35	26	-39	57	52	19	-41
21.	17	30	17	00	64	84	-43
22.	10	18	20	03	30	37	-56
23.	24	-05	25	-16	00	39	43
24.	-28	06	-28	37	-11	-71	-45
25.	23	00	20	-13	-09	15	36

more closed the school climate is, the smaller the percentage of teachers who left teaching or school.

The Ideology *lv* affects turnover rate mostly through the open vs. closed climate and affects satisfaction indirectly via the supportive climate, as well as directly. The more traditional the educational ideology is in the school, the more closed is the climate ($p = -.38$), and a closed climate supports teachers' perseverance. In addition, there is a direct effect of ideology on satisfaction. The more traditional the ideology, the greater the satisfaction of the teachers with their jobs, and the lower the percentage who change schools.

This mechanism is also supported by the indirect effect through the other dimension of the climate - the supportive one. The role the organizational structure plays in the suggested model is twofold: for one, the socio-cultural composition (SCC) of the school affects the supportive climate, and only via the supportive climate is the turnover rate affected by SCC. The second dimension of the Structure, the organizational components of it, hardly affect the supportive climate ($p = -.08$), but it does affect, to a larger extent, the open vs. closed climate ($p = -.41$).

In addition, the organizational structure affects the educational ideology to a large extent ($p = .72$), as was observed in Model I as well. The indirect effects of the organizational structure through the "psychological processes" are much greater than their direct effect on it. Thus, although the organizational structure is fairly correlated with both perseverance factors (see Table 2), educational ideology and climate dimensions serve as mediating factors in the "life cycle" of the school personnel.

The fit of the model to the data can be checked from Table 2. Both the explained variances and the squared multiple correlations are much higher in Model II than they were in the first model.

To summarize our analysis, it may be safely said that socio-psychological constructs within the school - the subjective, self-reported perceived educational ideology and school's climate do mediate between the structural components and teachers' retention. Taken as a group, these intervening processes account for a very large portion of the effect of the organizational structure on all perseverance components; the affective, cognitive and instrumental behaviors. About 75% of the effect of the structure on satisfaction and 94% of its effect on turnover rate are mediated by the ideology and climate dimensions. The important result is, then, that organizational structure, though

fairly correlated, has no direct influence on job-satisfaction and teachers' retention, if educational ideology and climate components are taken into consideration. Thus, prevailing beliefs that something can be done which is within the reach of the school administration in changing certain outcomes, receives support.

Since the results provide both structural relations and the relative importance of the indicators in determining them, it is possible to recheck the model with a new set of data. It is assumed, though, that replacement of indicator variables will not drastically alter the inner relations among the lv's even though their outer relations may be changed, provided the new sample will be from a similar population. That is, some generalizations can be drawn from the above analysis.

It also seems to us that the treatment of this set of data provides an example of possible alternative analyses for many evaluation problems, in which rich data and scarcity of knowledge are, by far, the more general case at hand. "When we think of influence on knowledge,... we rarely think of the influence of methodology" stated Kerlinger in his presidential address to the AERA in 1977, and he added: "this is strange because methodology has already had a profound influence on behavioral scientific knowledge. Methodological advances make a difference mainly because they change our ways of thinking about what we can study and how we can study it" (p. 10).

The writer believes that Herman Wold's approach may bring about an advancement in research thinking due largely to the methodological development. Problems that have seemed to be too complex to tackle, may now become tractable and amenable to scientific scrutiny.

REFERENCES

- DUNCAN, O.D. Path Analysis: Sociological Examples. *American Journal of Sociology*, 1966, 72, 1-16.
- HOPPLE, G.W. *Soft Modeling in Social Sciences: Application of Partial Least Squares in Political Science*. Virginia: International Public Policy Research Corporation, Publication No. 79-12, 1979.
- HOROWITZ, T. and ZAK, I. *Patterns of Recruitment and Constancy in the Teaching Profession. Part II: Teachers' Continuity and Mobility*. Jerusalem: The Szold Institute, Research Report No. 210, January 1979 (Hebrew with an English Summary).
- JORESKOG, K.G. Structural Analysis of Covariance and Correlation Matrices. *Psychometrika*, 1978, 43, 443-477.

- KERLINGER, F.N. Presidential Address to the American Educational Research Association, April 1977.
- LOHMOLLER, J.B. Estimation Parameters of Linear Structural Relation Models Under Partial Least-squares Criteria. A paper presented at AERA, April 1979.
- NOONAN, R. and WOLD, H. NIPLAS Path Modeling with Latent Variables: Analysing School Survey Data using Non Linear, Iterative Partial Least Squares. *Scandinavian Journal of Educational Research*, 1977, 21, 33-61.
- TUCKER, L.R. An Inter-battery Method of Factor Analysis. *Psychometrika*, 1958, 23, 111-136.
- VAN DEN WOLENBERG, A.L. Canonical Redundancy Analysis: An Alternative for Canonical Correlation Analysis. *Psychometrika*, 1977, 42, 207-219.
- WOLD, H. Estimation of Principal Components and Related Models by Iterative Least Squares. In: P.R. Krishnaiah (Ed.) *Multivariate Analysis*, New York: Academic Press, 1966.
- WOLD, H. Path Analysis with Latent Variables. In: L.M. Blalock et al. (Eds.) *Quantitative Sociology*. New York: Academic Press, 1975.
- WOLD, H. *Model Construction and Evaluation when Theoretical Knowledge is Scarce: An Example of the Use of Partial Least Squares*. Cahier 79.06 du Département D'Econometrique, Faculté des Sciences Economiques et Sociales, Université de Genève, 1979.
- WRIGHT, S. The Method of Path Analysis. *Annals of Mathematical Statistics*, 1934, 5, 162-215.
- ZAK, I. School Profiles. In: A. Lewy (Ed.) *The Evaluation Role*, 1981, 399-422.
- ZAK, I. and HOROWITZ, T. Developmental Transitions in the Professional Growth of Teachers. In: F.M. Ottobre (Ed.) *Assessing Teacher Effectiveness*. Princeton, N.J.: IAEA, 1980, 198-227.